



Научном већу Института за физику у Београду

Предлог за Годишњу награду за научни рад Института за физику у Београду

Поштовани,

Велико ми је задовољство да предложим др **Марију Митровић Данкулов**, научног сарадника, за Годишњу награду за научни рад Института за физику у Београду, за њен допринос разумевању различитих колективних феномена у социјалним системима, као и даљем развоју теорије комплексних мрежа.

Пошто се Годишња награда додељује за резултате из претходне две године, а колегиница Митровић Данкулов је од 2. децембра 2015. до 1. децембра 2016. године била на породичном одсуству, предлажем да се при одлучивању о додели награде, уместо периода од 2015. године, разматра период почев од 2014. године. У том периоду кандидаткиња објавила седам радова у међународним часописима категорије M21a и M21. У питању су публикације у изузетним часописима као што су *Nature*, *Nature Communications*, *Scientific Reports* и *PLOS One*:

1. Рад *Growing time lags threatens Nobel* је објављен у часопису *Nature* 2014. године. Привукао је велику пажњу светских медија, како оних посвећених науци (*Phys.org*, *Scientific America*), тако и оних који се баве општим темама (*USA Today*, *SPIEGEL ONLINE*, *Business Standard*). О пажњи коју је привукао рад говори и његов *Altmetric* индекс који га сврстава у 5% чланака који су привукли највећу пажњу икада.

2. Рад *Inferring human mobility using communication patterns* је објављен 2014. године у часопису *Scientific Reports*. У овом раду су раније развијени методи примењени на анализу и моделирање једне реалне техно-социјалне мреже (мреже мобилне телефоније). Овај модел може допринети развоју области као што су урбано планирање, планирање јавног превоза и епидемиологија. О значају рада говори и чињеница да већ има 15 цитата (извор *Web of Science*) које је добио од објављивања крајем августа 2014. године.

3. Рад *Quantifying randomness in real networks* је објављен 2015. године у часопису *Nature Communications* и представља значајан допринос у области теорије комплексних мрежа. У овом раду је по први пут одређен минималан скуп тополошких особина које одређују структуру реалне комплексне мреже и квантификовано колико се она разликује од случајних мрежа. Пошто се помоћу комплексних мрежа данас описују различити системи, физички, биолошки, технолошки и социјални, теорија комплексних мрежа постаје све важнија за истраживања у овим областима. Због тога је овај рад од великог значаја и за ове области, као и за примењене области које се на њих наслањају. О значају рада говори и чињеница да већ има 6 цитата (извор *Web of Science*) које је добио за релативно кратко време (рад је објављен крајем октобра 2015. године).



4. У раду *The dynamics of meaningful social interactions and the emergence of collective knowledge* објављеном 2015. године у часопису *Scientific Reports* су први пут употребљени методи статистичке физике и теорије комплексних мрежа за проучавање феномена колективног настанка знања у социјалним заједницама. Анализом комплексне бипартитне мреже којом се представљају интеркације између делова комплексне мреже истраживана је њихова кластеризација, док спектар снаге временске серије активности корисника показује да су оне карактерисане лавинама, сличним Бракхаузеновом шуму или лавинама у неким физичким системима као што су модели пешчаних лавина. Овај рад је значајан за физику комплексних система, а његови резултати ће бити значајни за примене у социологији и дисциплинама које се баве динамиком настанка знања и динамиком учења у групама. Рад има 5 цитата (извор Web of Science) које је добио од објављивања у јулу 2015. године.

5. Рад *Topology of innovation spaces in the knowledge networks emerging through questions-and-answers* је објављен 2016. године у часопису *PLOS One*. Претходно развијене методе алгебарске топологије графова су овде примењене за испитивање структура вишег реда у комбинаторном простору колективног знања. Методологија примењена у овом раду се може искористити за анализу шире класе система колективног настанка новог знања.

6. Рад *A theoretical model for the associative nature of conference participation* је објављен 2016. године у часопису *PLOS One*. У овом раду је по први пут квантификована динамика поновљених учешћа на научним конференцијама и дат теоријски модел за њихово описивање. Идентификовани социјални механизми који имају кључни значај за поновљена учешћа у активностима једне научне заједнице представљају основне факторе који одређују универзалну динамику социјалних група.

7. У раду *Associative nature of event participation dynamics: A network theory approach*, објављеном 2017. године у часопису *PLOS One*, примењена је теорија комплексних мрежа за истраживање еволуције социјалних мрежа у оквиру *Meetup* окружења. Резултати овог рада су показали да је универзална динамика поновљених учешћа у активностима групе највише зависи од повећања везивног социјалног капитала. Овај и претходни рад на значајан начин доприносе области управљања формалним и неформалним друштвеним групама и могу да помогне у планирању и унапређивању њиховог рада.

На основу описаних резултата колегинице Митровић Данкулов јасно је да њен досадашњи рад представља значајан допринос развоју физике комплексних система, као и неким другим научним областима, што је доказ интердисциплинарног карактера њеног истраживања.

Колегиница Митровић Данкулов има широку научну сарадњу са групама из Словеније, Италије, Индије, Израела и Финске. Ментор је на докторским студијама једној студенткињи чија одбрана докторске тезе се очекује до краја ове године. Колегиница Митровић Данкулов је била члан програмског комитета највеће и најзначајније конференције у области физике социјалних система (*International Conference on Computational Social Science*) од почетка њеног одржавања, а тренутно је у програмским комитетима већег броја међународних конференција из области комплексних система. Поред научног рада, колегиница је ангажована и у Иновационом центру Института за физику у Београду као заменик руководиоца.



Колегиница Митровић Данкулов је у својој научној каријери објавила 18 радова од којих је већина у међународним часописима категорије M21a и M21, као и једно поглавље у књизи, а до сада је одржала четири предавања по позиву на међународним научним скуповима. Према бази Web of Science радови др Марије Митровић Данкулов цитирани су 180 пута (без аутоцитата), а њен h-фактор је 9.

Имајући све наведено у виду, са задовољством предлажем др Марију Митровић Данкулов за Годишњу награду за научни рад Института за физику у Београду.

У Београду, 15. 03. 2017. године

др Александар Белић
научни саветник
Институт за физику у Београду

МАРИЈА МИТРОВИЋ ДАНКУЛОВ

Афилијација: Лабораторија за примену рачунара у науци,
Центар за изучавање комплексних система
Институт за физику у Београду;
Иновациони центар, Институт за физику у Београду

Датум и место рођења: 7. мај 1981. године, Ћуприја, Србија

Брачно стање: удата, једна ћерка



ПРОФИЛ

Енергична, амбициозна и вредна особа коју карактеришу одличне организационе и комуникационе вештине. Поседује богато знање и искуство у области теоријске физике и у програмирању. Главна област њеног интересовања и истраживања је статистичка физика комплексних система и теорија комплексних мрежа, а посебно физика колективних социјалних феномена.

Ауторка је 18 публикација у водећим часописима, једног поглавља у књизи и 30 предавања по позиву и саопштења на међународним научним скуповима. Број цитата на основу базе Web of Science је 217 (180 без аутоцитата), h-индекс 9, док је број цитата на основу Google scholar 455, h-индекс 12. Пуна листа публикација дата је у прилогу.

ОБРАЗОВАЊЕ

- 2000-2005 **Физички факултет**, Универзитет у Београду
дипломирани физичар (2005).
- 2005-2012 **Физички факултет**, Универзитет у Београду
Постдипломске студије на одсеку Физика кондензованог стања материје.
Дипломе: **магистратура** (2010)
Магистарска теза: *Налажење отежињених подструктура и неким реалним и компјутерски генерисаним комплексних мрежама*
докторат (2012)
Докторска теза: *Структура и динамика техносоцијалних мрежа*

РАДНО ИСКУСТВО

- 2005-2009 **истраживач приправник** у Лабораторији за примену рачунара у науци, Институт за физику у Београду
- 2009-2012 **истраживач приправник** на Одсеку за теоријску физику, Институт Јожеф Штефан, Љубљана, Словенија

2012-2014 **постдокторско усавршавање** на Одсеку за биомедицински инжењеринг и рачунарске науке, Аалто Универзитет, Еспо, Финска

2014-до сада **научни сарадник** у Лабораторији за примену рачунара у науци, Национални центар изузетних вредности за изучавање комплексних система, Институт за физику у Београду

2014-до сада **заменик руководиоца** Иновационог центра Института за физику у Београду

ПРОЈЕКТИ

2006-2009 **CX-CMCS: EU Centre of Excellence for Computer Modeling of Complex Systems**, ЕУ ФП6 пројект

2006-2009 **ОИ141035: Моделирање и нумеричке симулације комплексних физичких система**, Министарство науке Републике Србије, пројекти основних истраживања

2009-2006 **CYBEREMOTIONS-Collective Emotions in Cyberspace**, ЕУ ФП7 пројекат

2014-до сада **ОИ171017: Моделирање и нумеричке симулације многочестичних система** Министарство просвете, науке и технолошког развоја Републике Србије, пројекти основних истраживања

2014-до сада **COST Action TU1305: Social Networks and Travel Behavior**, члан менаџмент комитета

2015-до сада **VI-SEEM**, ЕУ Х2020 пројекат

2017-до сада **Upscaling Teslagram® technology based on variable and complex biological structures for security printing**, Програм сарадње науке и привреде, Фонд за иновациону делатност Србије

НАСТАВНО И ПЕДАГОШКО ИСКУСТВО

2013-до сада ментор на докторским студијама (Јелена Смиљанић, Електротехнички факултет, Универзитет у Београду)

2015 ментор на студентској пракси (Петар Тадић, Физички факултет, Универзитет у Београду)

2008 асистент на Advanced School in High Performance and GRID Computing, ICTP, Trieste, Italy

2007 асистент на Advanced School in High Performance Computing Tools for e-Science-joint DEMOCRITOS/INFM-eLab/SISSA-ICTP activity, ICTP, Trieste, Italy

ОДБОРИ, УРЕДНИШТВО, ПРОГРАМСКИ КОМИТЕТИ

2009-до сада редовно рефереише радове у часописима **PLoS One, Journal of Statistical Mechanics (JSTAT), European Physical Journal B, Frontiers, Scientific Reports, Applied Network Science, Computational Social Networks.**

- 2014 **члан програмског комитета** конференције *6th International Conference on Information Technologies and Information Society ITIS 2014*, 5.-7. новембар, Шмајершке Топлице, Словенија
- 2015-до сада **члан уредништва за интердисциплинарну физику** часопис *Frontiers*
- 2015 **члан програмског комитета** конференције *1st Annual International Conference on Computational Social Science 2015*, 8.-11. јун 2016, Хелсинки, Финска
- 2015 **члан програмског комитета** конференције *7th International Conference on Information Technologies and Information Society ITIS2015*, 4.-6. новембар 2015, Ново Место, Словенија
- 2016 **члан програмског комитета** конференције *2nd Annual International Conference on Computational Social Science 2016*, 23.-26. јун 2016, Еванстон, Илиноис, САД
- 2016 **члан програмског комитета** конференције *3rd Conference on Sustainable Urban Mobility – 3rd CSUM 2016*, 26.-27. мај, Волос, Грчка
- 2016 **члан програмског комитета** конференције *5th International Conference on Complex Networks and their Applications – COMPLEX NETWORKS 2016*, 30. новембар – 2. децембар 2016, Милано, Италија
- 2017 **члан програмског комитета** конференције *Conference on Complex Systems 2017*, 17.-22. септембар 2017, Канкун, Мексико
- 2017 **члан програмског комитета** конференције *6th International Conference on Complex Networks and their Applications – COMPLEX NETWORKS 2017*, 29. новембар – 1. децембар 2017, Лион, Француска
- 2017 **члан програмског комитета** конференције *2nd International Conference on Complexity, Future Information Systems and Risk” - COMPLEXIS 2017*, 24.-26. април 2017, Порто, Португал.
- 2017-до сада **члан** Одбора Међууниверзитетског програма за истраживање одрживог развоја Универзитета у Београду
- 2017 **уредник** *Frontiers Research Topic "Culturomics: Interdisciplinary Path Towards Quantitative Study of Human Culture"* часопис *Frontiers*

ОРГАНИЗАЦИЈА

- 2006 **Организатор** прве Студентске недеље у Београду, ISWiB, Србија, 30. 6. - 5. 7. 2006
- 2011 **Организатор** конференције *Cyberemotions – collective emotions in cyberspace*, Љубљана, Словенија, 20.-21. јануар 2011
- 2013 **Организатор** конференције *First annual meeting of COST Action TD1210 KNOWeSCAPE - Analyzing the dynamics of information and knowledge landscapes*, Еспо, Финска, 18-20 новембар 2013

МЕЂУНАРОДНА САРАДЊА

2006-до сада	Одсек за теоријску физику, Институт Јожеф Штефан, Љубљана, Словенија професор Босиљка Тадић
2012-до сада	Одсек за биомедицински инжењеринг и рачунарске науке, Аалто Универзитет, Еспо, Финска, професор Санто Фортунато
2012-до сада	Одсек за физику чврстог стања, Саха институт за нуклеарне науке, др Арнаб Чатерџи
2014-до сада	Аалто Универзитет, Еспо, Финска, др Томи Каупинен
2015-до сада	Универзитет Технион, Хаифа, Израел, проф. др Пнина Плаут
2015-до сада	Универзитет Милано Бикока, Милано, Италија, проф. др Силвана Стефани

НАГРАДЕ И СТИПЕНДИЈЕ

2005-2006	Министарство науке, Република Србија
2004	Влада Краљевине Норвешке

ЈЕЗИЦИ

српски-матерњи; енглески - одлично; словеначки - основни ниво;

ПРОГРАМИРАЊЕ И ПОЗНАВАЊЕ РАДА НА РАЧУНАРУ

програмирање: C/C++, Pascal, Matlab, Mathematica, Python, shell scripts

напредни корисник Linux, Mac OS и Windows оперативних система, Grid computing

ОСТАЛЕ АКТИВНОСТИ

свира хармонику, завршила је Нижу музичку школу “Душан Сковран” у Ћуприји, смер хармоника

ауторка је и научно-популарних чланка у часописима Млади физичар (Друштво физичара Србије), Wavemagazine (www.wavemagazine.net) и Беседе (Друштво српска заједница, Љубљана, Словенија)

глумила је у представи *Зашто пишем песме* (Друштво српска заједница, Љубљана, Словенија)

Листа публикација др Марије Митровић Данкулов за релевантан период за награду у часописима категорије M20

1. *Associative nature of event participation dynamics: A network theory approach*
J. Smiljanić and **M. Mitrović Dankulov**
PLoS ONE **12**, e0171565 (2017)
Категорија: M21; ИФ=3.057 (подаци за 2015. годину)
2. *Topology of Innovation Spaces in the Knowledge Networks Emerging through Questions-And-Answers*
M. Andjelković, B. Tadić, **M. Mitrović Dankulov**, M. Rajković, and R. Melnik
PLoS ONE **11**, e0154655 (2016)
Категорија: M21; ИФ=3.057 (подаци за 2015. годину)
3. *A Theoretical Model for the Associative Nature of Conference Participation*
J. Smiljanić, A. Chatterjee, T. Kauppinen, and **M. Mitrović Dankulov**
PLoS ONE **11**, e0148528 (2016)
Категорија: M21; ИФ=3.057 (подаци за 2015. годину)
4. *Quantifying Randomness in Real Networks*
C. Orsini, **M. Mitrović Dankulov**, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov
Nat. Commun. **6**, 8627 (2015)
Категорија: M21a; ИФ=11.329 (подаци за 2015. годину)
5. *The Dynamics of Meaningful Social Interactions and the Emergence of Collective Knowledge*
M. Mitrović Dankulov, R. Melnik, and B. Tadić
Sci. Rep. **5**, 12197 (2015)
Категорија: M21; ИФ=5.228 (подаци за 2015. годину)
6. *Inferring Human Mobility Using Communication Patterns*
V. Palchykov, **M. Mitrović**, H. Jo, J. Saramaki, and R. Ku. Pan
Sci. Rep. **4**, 6174 (2014)
Категорија: M21a; ИФ=5.578 (подаци за 2014. годину)
7. *Growing Time Lag Threatens Nobels*
S. Fortunato, A. Chatterjee, **M. Mitrović**, R. Ku. Pan, P. Della Briotta Parolo, and F. Becattini
Nature **508**, 186 (2014)
Категорија=M21a; ИФ=41.456 (подаци за 2014. годину)

Листа публикација др Марије Митровић Данкулов

Публикације М20

1. *Associative nature of event participation dynamics: A network theory approach*
J. Smiljanić and **M. Mitrović Dankulov**
PLoS ONE **12**, e0171565 (2017)
Категорија: М21; ИФ=3.057 (подаци за 2015. годину)
2. *Topology of Innovation Spaces in the Knowledge Networks Emerging through Questions-And-Answers*
M. Andjelković, B. Tadić, **M. Mitrović Dankulov**, M. Rajković, and R. Melnik
PLoS ONE **11**, e0154655 (2016)
Категорија: М21; ИФ=3.057 (подаци за 2015. годину)
3. *A Theoretical Model for the Associative Nature of Conference Participation*
J. Smiljanić, A. Chatterjee, T. Kauppinen, and **M. Mitrović Dankulov**
PLoS ONE **11**, e0148528 (2016)
Категорија: М21; ИФ=3.057 (подаци за 2015. годину)
4. *Quantifying Randomness in Real Networks*
C. Orsini, **M. Mitrović Dankulov**, P. Colomer-de-Simón, A. Jamakovic, P. Mahadevan, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli, S. Fortunato, and D. Krioukov
Nat. Commun. **6**, 8627 (2015)
Категорија: М21а; ИФ=11.329 (подаци за 2015. годину)
5. *The Dynamics of Meaningful Social Interactions and the Emergence of Collective Knowledge*
M. Mitrović Dankulov, R. Melnik, and B. Tadić
Sci. Rep. **5**, 12197 (2015)
Категорија: М21; ИФ=5.228 (подаци за 2015. годину)
6. *Inferring Human Mobility Using Communication Patterns*
V. Palchykov, **M. Mitrović**, H. Jo, J. Saramaki, and R. Ku. Pan
Sci. Rep. **4**, 6174 (2014)
Категорија: М21а; ИФ=5.578 (подаци за 2014. годину)
7. *Growing Time Lag Threatens Nobels*
S. Fortunato, A. Chatterjee, **M. Mitrović**, R. Ku. Pan, P. Della Briotta Parolo, and F. Becattini
Nature **508**, 186 (2014)
Категорија=М21а; ИФ=41.456 (подаци за 2014. годину)

8. *Co-Evolutionary Mechanisms of Emotional Bursts in Online Social Dynamics and Networks*
B. Tadić, V. Gligorijević, **M. Mitrović**, and M. Šuvakov
Entropy **15**, 5084 (2013)
Категорија=M22; ИФ=1.564 (подаци за 2013. годину)
9. *How the online social networks are used: dialogues-based structure of MySpace*
M. Šuvakov, **M. Mitrović**, V. Gligorijević, and B. Tadić
J. R. Soc. Interface **10**, 20120819 (2013)
Категорија=M21a; ИФ=4.907 (подаци за 2013. годину)
10. *Universality in voting behavior: an empirical analysis*
A. Chatterjee, **M. Mitrović**, and S. Fortunato
Sci. Rep. **3**, 1049 (2013)
Категорија: M21a; ИФ=5.078 (подаци за 2013. годину)
11. *Statistical Analysis of Emotions and Opinions at Digg Website*
P. Pohorecki, J. Sienkiewicz, **M. Mitrović**, G. Paltoglou, and J. A. Holyst
Acta Phys. Pol. A **123**, 604 (2013)
Категорија: M23; ИФ=0.604 (подаци за 2013. годину)
12. *Dynamics of bloggers' communities: Bipartite networks from empirical data and agent-based modeling*
M. Mitrović and B. Tadić
Physica A **391**, 5264 (2012)
Категорија: M22; ИФ=1.676 (подаци за 2012. годину)
13. *Quantitative analysis of bloggers' collective behavior powered by emotions*
M. Mitrović, G. Paltoglou, and B. Tadić
J. Stat. Mech.-Theory Exp. P02005 (2011)
Категорија: M21; ИФ=1.727 (подаци за 2011. годину)
14. *Network theory approach for data evaluation in the dynamic force spectroscopy of biomolecular interactions*
J. Živković, **M. Mitrović**, L. Janssen, H. A. Heus, B. Tadić, and S. Speller
EPL **89**, 68004 (2010)
Категорија: M21; ИФ=2.753 (подаци за 2010. годину)
15. *Bloggers behavior and emergent communities in Blog space*
M. Mitrović and B. Tadić
Eur. Phys. J. B **73**, 293 (2010)
Категорија: M22; ИФ=1.575 (подаци за 2010. годину)

16. *Networks and emotion-driven user communities at popular Blogs*
M. Mitrović, G. Paltoglou, and B. Tadić
Eur. Phys. J. B **77**, 597 (2010)
Категорија: M22; ИФ=1.575 (подаци за 2010. годину)
17. *Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities*
M. Mitrović and B. Tadić
Phys. Rev. E **80**, 026123 (2009)
Категорија: M21; ИФ=2.400 (подаци за 2009. годину)
18. *Jamming and correlation patterns in traffic of information on sparse modular networks*
B. Tadić and **M. Mitrović**
Eur. Phys. J. B **71**, 631 (2009)
Категорија: M22; ИФ=1.466 (подаци за 2009. годину)

Публикације M10

1. *Emergence and structure of cybercommunities*
M. Mitrović and B. Tadić
In *Springer Handbook of Optimization in Complex Networks Theory and Applications, part 2: "Structure and Dynamics of Complex Networks"*
Ed. M. M. Thai and P. Pardalos, 57, Part 2, pp. 209-227, Springer, Berlin (2012)
Категорија: M13

Публикације M30

Предавања по позиву:

1. *Quantifying and Modeling Collective Behavior in (online) Social Systems*,
M. Mitrović Dankulov, Winter Workshop on Complex Systems 2017
(WWCS 2017), February 6-10, 2017, Petnica, Serbia. Категорија: M32
2. *Complex Networks Theory: An Introduction*,
M. Mitrović Dankulov, Summer School on Topological and Scaling
Analysis of Transport and Social Media Data, June 13-17, 2016, Gavle,
Sweden. Категорија: M32

3. *Quantitative Study and Modeling of Collective Knowledge Building via Questions and Answers*, **M. Mitrović Dankulov** and B. Tadić, 19th Symposium on Condensed Matter Physics, SFKM2015, September 7-11, 2015, Belgrade, Serbia. Категорија: M32
4. *Agent-Based Modeling and Social Structure in Bloggers' Dynamics*, **M. Mitrović** and B. Tadić, 6th Summer Solstice International Conference on Discrete Models of Complex Systems, SUMMERSOLSTICE 2014, June 22-25, 2014, Ljubljana, Slovenia. Категорија: M32

Остала предавања и саопштења:

5. *Correlation patterns in gene expressions along the cell cycle of yeast* J. Živković, **M. Mitrović** and B. Tadić, Proceedings of Complex Networks: results of the 2009 International Workshop on Complex Networks (CompleNet 2009), [26-27 May 2009, Catania, Italy], Studies in computational intelligence, 207, Berlin; Heiderberg: Springer, 23-33 (2009). Категорија: M34
6. *Mixing patterns and communities on bipartite graphs on web-based social interactions*, J. Grujić, **M. Mitrović** and B. Tadić, Proceedings of 16th International Conference on Digital Signal Processing, July 5-7 2009, Santorini, Greece. DSP 2009. New York: IEEE, 1-8, (2009). Категорија: M33
7. *Congestion patters of traffic studied on Nnjing city dual graph*, H.-L. Zeng, Y.-D. Guo, C.-P. Zhu, **M. Mitrović** and B. Tadić, Proceedings of 16th International Conference on Digital Signal Processing, July 5-7 2009, Santorini, Greece, DSP 2009. New York: IEEE, pp.1-8, (2009). Категорија: M33
8. *Search of weighted subgraphs on complex networks with maximum likelihood methods*, **M. Mitrović** and B. Tadić, LNCS, part 2, INCS 5102, pp. 551-558 (2008). . Категорија: M33
9. *Conference attendance patterns*, J. Smiljanić and **M. Mitrović Dankulov**, 19th Symposium on Condensed Matter Physics, SFKM2015, September 7-11, 2015, Belgrade, Serbia. Категорија: M34

10. *Algebraic Topology Analysis of Networks Emerging from Content-Driven Social Interactions*, M. Andjelković, B. Tadić, **M. Mitrović**, and M. Rajković, From Data to Knowledge, the Third Annual Knowscape Conference, October 7-9, 2015, Mons, Belgium. Категорија: M34
11. *The dynamics of collective knowledge building via questions and answers*, **M. Mitrović Dankulov** and B. Tadić, International Conference on Computational Social Science, June 8-11, 2015, Helsinki, Finland. Категорија: M34
12. *Modeling The Dynamics of Knowledge Creation in Online Communities*, B. Tadić and **M. Mitrović Dankulov**, 7th International Conference on Discrete Models of Complex Systems, 2015 Summer Solstice, June 17-19, 2015, Toronto, Canada. . Категорија: M34
13. *Quantitative Study of Innovation and Knowledge Building in Questions&Answers System with Math Tags*, **M. Mitrović** and B. Tadić , The Second Annual KnowEscape Conference, KnowEscape2014, November 24-26, 2014, Thessaloniki, Greece. Категорија: M34
14. *The Death of Expertise & Problems in Quantifying Collective Knowledge in Online Social*, B. Tadić and **M. Mitrović**, The Second Annual KnowEscape Conference, KnowEscape2014, Thessaloniki, Greece, November 24-26, 2014. Категорија: M34
15. *Universality in Voting Behavior*, **M. Mitrović**, A. Chatterjee and S. Fortunato, The Second National Conference Information Theory and Complex Systems, TINKOS 2014, Niš, Serbia, June 16-17, 2014. Категорија: M34
16. *Universal Patterns of Voting Behavior*, **M. Mitrović**, A. Chatterjee and S. Fortunato, The Second Annual KnowEscape Conference, KnowEscape2014, Helsinki, Finland, November 18-20, 2013. Категорија: M34
17. *Agent-Based Model Of Blogging*, **M. Mitrović** and B. Tadić, European Conference on Complex Systems, Brussels, Belgium, September 3-7, 2012. Категорија: M34
18. *Network-based methodology for analysis of complex systems: theory & applications*, **M. Mitrović** and B. Tadić XVII National Symposium on Condensed Matter Physics, SFKM 2011, Belgrade, Serbia, April 18-22, 2011. Категорија: M34

19. *Complexity in the dynamics of Web users: Methodology for quantitative analysis of empirical data and simulations*, **M. Mitrović** and B. Tadić, European Conference on Complex Systems, Vienna, Austria, September 12- 16, 2011. Категорија: M34
20. *Modeling of emotional agents on Blogs*, **M. Mitrović** and B. Tadić, Cyberemotions - collective emotions in cyberspace, Ljubljana, Slovenia, 20-21 September, 2011. Категорија: M34
21. *Network based methodology for analysis of on-line collective behavior* **M. Mitrović** COST action NP0801 Third Annual Meeting: Physics of Competition and Conicts, Eindhoven, Netherlands, May 18-20, 2011. Категорија: M34
22. *Bipartite network analysis reveals the role of emotion in comments on digg stories*, **M. Mitrović**, Processes on networks: hunting for universality in social, economical and Biological Networks, COST Woskhop, 10-12 March 2010, Vienna, Austria, 2010. Категорија: M34
23. *Emotions & user communities in Blogs and Diggs*, **M. Mitrović** and B. Tadić, The CyberEmotions Workshop, 21-23 January, Wolverhampton, UK, 2010. Категорија: M34
24. *Network structure and emotions on popular posts*, **M. Mitrović** and B. Tadić, COST action NP0801 Second Annual Meeting: Physics of Competition and Conflicts, Sunny Beach, Bulgaria, May 26-28, 2010. Категорија: M34
25. *Patterns of user behavior and community structure on blogs*, **M. Mitrović** and B. Tadić, TWCS 2010, Turunc Workshop on Complex, 30 August - 1 September 2010, Turunc, Marmaris Turkey, 2010. Категорија: M34
26. *Agent based model for use behaviour on emergent networks*, **M. Mitrović** and B. Tadić, Cyberemotions - collective emotions in cyberspace, Lousanne, 8-9 September 2010. Категорија: M34
27. *Collective emotional behavior on blogs : data-driven modeling and theoretical survey*, B. Tadić, **M. Mitrović** and G. Paltoglou, Proceedings of ECCS'10 Lisbon, European Conference on Complex Systems'10, September 13-17, 2010, Lisbon, Portugal. Категорија: M34

28. *Spectral analysis of networks reveals communities in complex systems data*, **M. Mitrović** and B. Tadić, COST action NP0801 First Annual Meeting: Physics of Competition and Conicts and NET 2009: evolution and complexity, Rome, May 28th-30th, 2009. Категорија: M34
29. *Finding structure in Blogs: bipartite networks analysis: invited presentation, extended abstract*, **M. Mitrović** and B. Tadić, Proceeding VALUETOOLS '09, Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools, 2009. Категорија: M34
30. *Modularity of networks from the perspective of spectral analysis*, **M. Mitrović**, International Workshop and Seminar on “Bio-inspired complex networks in Science and Technology”, Max Planck Institute for the Physics of Complex Systems in Dresden, Germany, 2008. Категорија: M34

Citation Report: 22

(from Web of Science Core Collection)

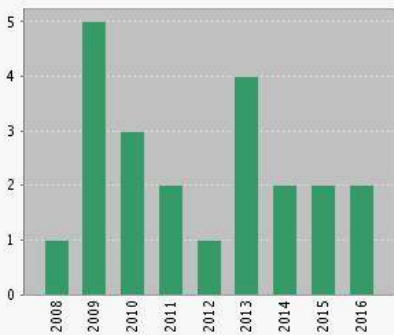
You searched for: **AUTHOR IDENTIFIERS:** (I-3007-2012)

Timespan: All years. Indexes: SCI-EXPANDED, SSCI, A&HCI, CPCI-S, CPCI-SSH, ESCI.

[...Less](#)

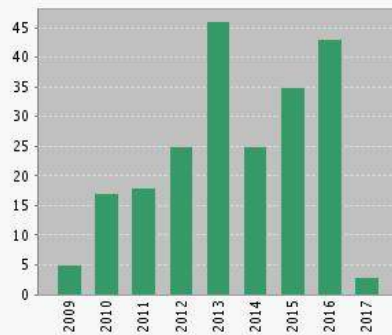
This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.

Published Items in Each Year



The latest 20 years are displayed.

Citations in Each Year



The latest 20 years are displayed.

Results found:	22
Sum of the Times Cited [?]:	217
Sum of Times Cited without self-citations [?]:	180
Citing Articles [?]:	157
Citing Articles without self-citations [?]:	145
Average Citations per Item [?]:	9.86
h-index [?]:	9

Sort by: Times Cited -- highest to lowest

Page 1 of 3

	2013	2014	2015	2016	2017	Total	Average Citations per Year
Use the checkboxes to remove individual items from this Citation Report or restrict to items published between 1996 and 2017 Go	46	25	35	43	3	217	24.11
1. Spectral and dynamical properties in classes of sparse networks with mesoscopic inhomogeneities By: Mitrovic, Marija; Tadic, Bosiljka PHYSICAL REVIEW E Volume: 80 Issue: 2 Article Number: 026123 Published: AUG 2009	5	7	5	7	0	37	4.11
2. Networks and emotion-driven user communities at popular blogs By: Mitrovic, M.; Paltoglou, G.; Tadic, B. EUROPEAN PHYSICAL JOURNAL B Volume: 77 Issue: 4 Pages: 597-609 Published: OCT 2010	11	4	1	0	0	31	3.88
3. Bloggers behavior and emergent communities in Blog space By: Mitrovic, M.; Tadic, B. EUROPEAN PHYSICAL JOURNAL B Volume: 73 Issue: 2 Pages: 293-301 Published: JAN 2010	6	2	1	1	0	25	3.12
4. Universality in voting behavior: an empirical analysis By: Chatterjee, Arnab; Mitrovic, Marija; Fortunato, Santo SCIENTIFIC REPORTS Volume: 3 Article Number: 1049 Published: JAN 10 2013	7	2	4	5	1	19	3.80
5. Quantitative analysis of bloggers' collective behavior powered by emotions By: Mitrovic, Marija; Paltoglou, Georgios; Tadic, Bosiljka JOURNAL OF STATISTICAL MECHANICS-THEORY AND EXPERIMENT Article Number: P02005 Published: FEB 2011	7	0	4	0	0	16	2.29

6. **Inferring human mobility using communication patterns**
 By: Palchykov, Vasyli; Mitrovic, Marija; Jo, Hang-Hyun; et al.
[SCIENTIFIC REPORTS](#) Volume: 4 Article Number: 6174 Published: AUG 22 2014

7. **Co-Evolutionary Mechanisms of Emotional Bursts in Online Social Dynamics and Networks**
 By: Tadic, Bosiljka; Gligorijevic, Vladimir; Mitrovic, Marija; et al.
[ENTROPY](#) Volume: 15 Issue: 12 Pages: 5084-5120 Published: DEC 2013

8. **How the online social networks are used: dialogues-based structure of MySpace**
 By: Suvakov, Milovan; Mitrovic, Marija; Gligorijevic, Vladimir; et al.
[JOURNAL OF THE ROYAL SOCIETY INTERFACE](#) Volume: 10 Issue: 79 Article Number: 20120819 Published: FEB 6 2013

9. **Dynamics of bloggers' communities: Bipartite networks from empirical data and agent-based modeling**
 By: Mitrovic, Marija; Tadic, Bosiljka
[PHYSICA A-STATISTICAL MECHANICS AND ITS APPLICATIONS](#) Volume: 391 Issue: 21 Pages: 5264-5278 Published: NOV 1 2012

10. **Jamming and correlation patterns in traffic of information on sparse modular networks**
 By: Tadic, B.; Mitrovic, M.
[EUROPEAN PHYSICAL JOURNAL B](#) Volume: 71 Issue: 4 Pages: 631-640 Published: OCT 2009

0	1	7	6	1	15	3.75
0	4	4	3	0	11	2.20
2	1	5	3	0	11	2.20
4	3	1	2	0	10	1.67
2	0	0	1	0	8	0.89

Select Page



Save to Text File



Sort by: **Times Cited -- highest to lowest** ▼

◀ Page 1 of 3 ▶

22 records matched your query of the 36,843,663 in the data limits you selected.

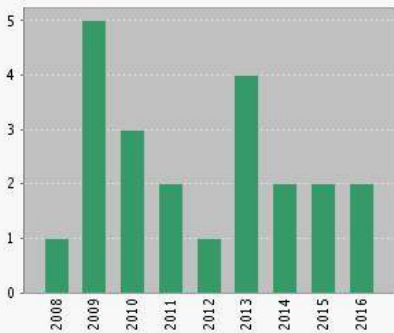
Citation Report: 22

(from Web of Science Core Collection)

You searched for: **AUTHOR IDENTIFIERS:** (I-3007-2012) ...More

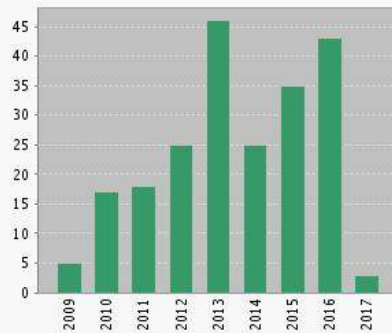
This report reflects citations to source items indexed within Web of Science Core Collection. Perform a Cited Reference Search to include citations to items not indexed within Web of Science Core Collection.

Published Items in Each Year



The latest 20 years are displayed.

Citations in Each Year



The latest 20 years are displayed.

Results found:	22
Sum of the Times Cited [?]:	217
Sum of Times Cited without self-citations [?]:	180
Citing Articles [?]:	157
Citing Articles without self-citations [?]:	145
Average Citations per Item [?]:	9.86
h-index [?]:	9

Sort by: Times Cited -- highest to lowest

Page 2 of 3

	2013	2014	2015	2016	2017	Total	Average Citations per Year
Use the checkboxes to remove individual items from this Citation Report or restrict to items published between 1996 and 2017 Go	46	25	35	43	3	217	24.11
11. Search of weighted subgraphs on complex networks with maximum likelihood methods By: Mitrovic, Marija; Tadic, Bosiljka Edited by: Bubak, M; VanAlbada, GD; Dongarra, J; et al. Conference: 8th International Conference on Computational Science Location: Cracow, POLAND Date: JUN 23-25, 2008 Sponsor(s): Hewlett Packard Co; Intel Corp; Qumak Sekom AM & IBM; Microsoft Corp; ATM SA; Elsevier; Springer COMPUTATIONAL SCIENCE - ICCS 2008, PT 2 Book Series: Lecture Notes in Computer Science Volume: 5102 Pages: 551+ Published: 2008	0	0	0	0	0	8	0.80
12. Quantifying randomness in real networks By: Orsini, Chiara; Dankulov, Marija M.; Colomer-de-Simon, Pol; et al. NATURE COMMUNICATIONS Volume: 6 Article Number: 8627 Published: OCT 2015	0	0	0	5	1	6	2.00
13. The dynamics of meaningful social interactions and the emergence of collective knowledge By: Dankulov, Marija Mitrovic; Melnik, Roderick; Tadic, Bosiljka SCIENTIFIC REPORTS Volume: 5 Article Number: 12197 Published: JUL 15 2015	0	0	0	5	0	5	1.67
14. Correlation Patterns in Gene Expressions along the Cell Cycle of Yeast By: Zivkovic, Jelena; Mitrovic, Marija; Tadic, Bosiljka Edited by: Fortunato, S; Mangioni, G; Menezes, R; et al. Conference: International Workshop on Complex Networks (CompleNet 2009) Location: Univ Catania, Catania, ITALY Date: MAY 26-27, 2009 COMPLEX NETWORKS Book Series: Studies in Computational Intelligence Volume: 207 Pages: 23+ Published: 2009	0	0	0	1	0	5	0.56

- 15. **Statistical Analysis of Emotions and Opinions at Digg Website**
 By: Pohorecki, P.; Sienkiewicz, J.; Mitrovic, M.; et al.
 Conference: 6th Polish Symposium of Physics in Economy and Social Sciences (FENS) Location: Univ Gdansk, Fac Math, Phys & Informat, Gdansk, POLAND
 Date: APR 19-21, 2012
[ACTA PHYSICA POLONICA A](#) Volume: 123 Issue: 3 Pages: 604-614
 Published: MAR 2013
- 16. **Growing time lag threatens Nobels**
 By: Fortunato, Santo
[NATURE](#) Volume: 508 Issue: 7495 Pages: 186-186 Published: APR 10 2014
- 17. **CYBEREMOTIONS - Collective Emotions in Cyberspace**
 By: Ahn, Junghyun; Borowiec, Anna; Buckley, Kevan; et al.
 Edited by: Giacobino, E; Pfeifer, R
 Conference: 2nd European Future Technologies Conference and Exhibition (FET) Location: Budapest, HUNGARY Date: MAY 04-06, 2011
 Sponsor(s): European Commiss Future & Emerging Technol (FET); European Res Consortium Informat & Mathemat (ERCIM); Hungarian Acad Sci; Hungarian Presidency European Un
 PROCEEDINGS OF THE 2ND EUROPEAN FUTURE TECHNOLOGIES CONFERENCE AND EXHIBITION 2011 (FET 11) Book Series: Procedia Computer Science Volume: 7 Pages: 221-+ Published: 2011
- 18. **Network theory approach for data evaluation in the dynamic force spectroscopy of biomolecular interactions**
 By: Zivkovic, J.; Mitrovic, M.; Janssen, L.; et al.
[EPL](#) Volume: 89 Issue: 6 Article Number: 68004 Published: MAR 2010
- 19. **Topology of Innovation Spaces in the Knowledge Networks Emerging through Questions-And-Answers**
 By: Andjelkovic, Miroslav; Tadic, Bosiljka; Dankulov, Marija Mitrovic; et al.
[PLOS ONE](#) Volume: 11 Issue: 5 Article Number: e0154655 Published: MAY 12 2016
- 20. **A Theoretical Model for the Associative Nature of Conference Participation**
 By: Smiljanic, Jelena; Chatterjee, Arnab; Kauppinen, Tomi; et al.
[PLOS ONE](#) Volume: 11 Issue: 2 Article Number: e0148528 Published: FEB 9 2016

1	1	2	0	0	4	0.80
0	0	1	2	0	3	0.75
0	0	0	2	0	2	0.29
1	0	0	0	0	1	0.12
0	0	0	0	0	0	0.00
0	0	0	0	0	0	0.00

Select Page   

Sort by:

◀ Page of 3 ▶

22 records matched your query of the 36,843,663 in the data limits you selected.

Growing time lag threatens Nobels

The time lag between reporting a scientific discovery worthy of a Nobel prize and the awarding of the medal has increased, with waits of more than 20 years becoming common. If this trend continues, some candidates might not live long enough to attend their Nobel ceremonies.

Before 1940, Nobels were awarded more than 20 years after the original discovery for only about 11% of physics, 15% of chemistry and 24% of physiology or medicine prizes, respectively. Since 1985, however, such lengthy delays have featured in 60%, 52% and 45% of these awards, respectively.

The increasing average interval between reporting discoveries and their formal recognition can be fitted to an exponential curve (see 'The long road to Sweden'), with data points scattered about the mean value.

As this average interval becomes longer, so the average age at which laureates are awarded the prize goes up. By the end of this century, the prizewinners' predicted average age for receiving the award is likely to exceed his or her projected life expectancy (data not shown). Given that the Nobel prize cannot be awarded posthumously, this lag threatens to undermine science's most venerable institution.

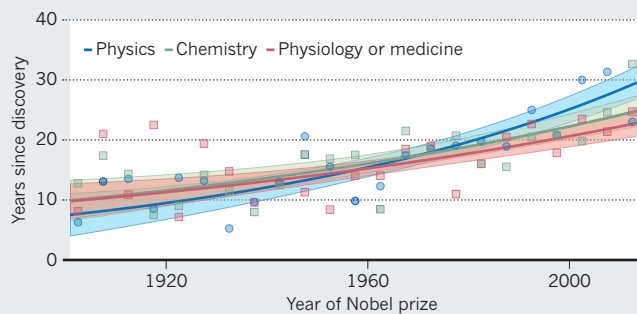
Santo Fortunato* *Aalto University, Finland.*
santo.fortunato@gmail.com
*On behalf of 6 co-authors; see go.nature.com/cmmxa5 for full list.

Livestock: tackle demand and yields

Among many otherwise laudable suggestions, Mark Eisler and colleagues propose limiting feedstuffs for livestock to fibrous fodder, such as grass and silage (see *Nature* 507, 32–34; 2014). However, we believe that any attempt to meet the rapid growth

THE LONG ROAD TO SWEDEN

Scientists who publish prizewinning discoveries are, on average, waiting longer for a Nobel than ever before.



Data points are 5-year-averaged waits; shading around lines shows confidence limits.

in world demand for meat and dairy products by focusing on ruminant grazing systems would be damaging for biodiversity and for the global climate.

Although ruminants convert grass and silage into animal protein, they do so inefficiently; they therefore require much more land to produce a given amount of meat or milk than ruminants fed on diets that include grain. Growing enough fodder to satisfy demand would require the large-scale expansion of grazing lands (see go.nature.com/7mf63y) — a leading cause of biodiversity loss, tropical deforestation and carbon dioxide emissions.

The environmental impacts of meat and dairy production should instead be addressed by stringent efforts to decrease consumption, halt the expansion of grazing, and increase yields on land that is already used for livestock. Promoting extensive grazing without tackling demand would do more harm than good.

Erasmus K. H. J. zu Ermgassen, *David R. Williams, Andrew Balmford* *University of Cambridge, UK.*
ekhjz2@cam.ac.uk

Livestock: limit red meat consumption

Mark Eisler and co-authors advocate eating only 300 grams of red meat a week (roughly the volume of three decks of

playing cards) as a step towards producing sustainable livestock (*Nature* 507, 32–34; 2014). That amount corresponds to 3.5–7% of a 2,000-calorie-a-day diet, depending on the cut and type of meat. Such a move would also make for a more equitable global distribution of animal-product consumption; these products comprise around 48% of the average diet in the United States, for example (S. Bonhommeau *et al. Proc. Natl Acad. Sci. USA* 110, 20617–20620; 2013).

Imposing a global dietary limit of 5% red meat as part of a 10% maximum for all animal-based products would enable more people to be fed using less land. For example, eliminating livestock and using existing agricultural lands to grow crops for direct human consumption instead of for livestock fodder could feed an extra 4 billion people (E. S. Cassidy *et al. Environ. Res. Lett.* 8, 034015; 2013), thereby reducing or eliminating the greenhouse-gas emissions and biodiversity loss associated with conversion of natural habitats. This would also reduce many other environmental impacts of agriculture that relate to the use of water, fertilizer and fossil fuels. **Brian Machovina, Kenneth J. Feeley** *Florida International University, Miami; and The Fairchild Tropical Botanic Garden, Coral Gables, Florida, USA.*
brianmachovina@gmail.com

Zoo visits boost biodiversity literacy

Zoos and aquaria worldwide attract more than 700 million visits every year. They are therefore well placed to make more people aware of the importance of biodiversity — a prime target of the United Nations Strategic Plan for Biodiversity 2011–20.

We surveyed approximately 6,000 visitors to 30 zoos and aquaria in 19 countries (see go.nature.com/vwf8yf). More respondents showed improved understanding of biodiversity after their visit (75.1% compared with 69.8% before) and more could identify an individual action that would bolster biodiversity after their visit (58.8% compared with 50.5% before).

Regrettably, increased awareness does not necessarily change behaviour. The world's zoo and aquarium communities must also help to drive important behavioural and social changes to assist conservation.

Andrew Moss *Chester Zoo, UK.*
Eric Jensen *University of Warwick, Coventry, UK.*
Markus Gusset *World Association of Zoos and Aquariums, Gland, Switzerland.*
markus.gusset@waza.org

A protein that spells trouble

The gene *CYLD* is so named because one of its mutant forms is associated with cylindromatosis, which causes skin tumours.

The *CYLD* protein is an enzyme; its active site in humans contains a cysteine residue at position 601 (denoted as C in the one-letter amino-acid code). The amino-acid sequence following this cysteine (C) is tyrosine (Y), leucine (L) and aspartate (D). What are the odds of that? **David Boone** *Indiana University School of Medicine — South Bend, Indiana, USA.*
daboone@iu.edu

**Supplementary information to:
Growing time lag threatens Nobels**

Full list of co-signatories to a Correspondence published in Nature **508**, 186 (2014);
<http://dx.doi.org/10.1038/508186a>

**Santo Fortunato, Arnab Chatterjee, Marija Mitrovic, Raj Kumar
Pan, Pietro Della Briotta Parolo** *Aalto University, Finland.*
santo.fortunato@gmail.com
Francesco Becattini *University of Florence and INFN Section of
Florence, Italy.*



OPEN

Inferring human mobility using communication patterns

SUBJECT AREAS:

APPLIED MATHEMATICS
COMPUTATIONAL SCIENCE
APPLIED PHYSICSReceived
1 May 2014Accepted
6 August 2014Published
22 August 2014Correspondence and
requests for materials
should be addressed to
R.K.P. (raj कुमार.pan@
aalto.fi)Vasyly Palchykov^{1,2,3}, Marija Mitrović^{1,4}, Hang-Hyun Jo^{1,5}, Jari Saramäki¹ & Raj Kumar Pan¹

¹Department of Biomedical Engineering and Computational Science (BECS), Aalto University School of Science, P.O. Box 12200, FI-00076, Finland, ²Institute for Condensed Matter Physics, National Academy of Sciences of Ukraine, UA 79011 Lviv, Ukraine, ³Lorentz Institute, Leiden University, 2300 RA Leiden, The Netherlands, ⁴Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Pregrevaica 118, 11080 Belgrade, Serbia, ⁵BK21plus Physics Division and Department of Physics, Pohang University of Science and Technology, Pohang 790-784, Republic of Korea.

Understanding the patterns of mobility of individuals is crucial for a number of reasons, from city planning to disaster management. There are two common ways of quantifying the amount of travel between locations: by direct observations that often involve privacy issues, e.g., tracking mobile phone locations, or by estimations from models. Typically, such models build on accurate knowledge of the population size at each location. However, when this information is not readily available, their applicability is rather limited. As mobile phones are ubiquitous, our aim is to investigate if mobility patterns can be inferred from aggregated mobile phone call data alone. Using data released by Orange for Ivory Coast, we show that human mobility is well predicted by a simple model based on the frequency of mobile phone calls between two locations and their geographical distance. We argue that the strength of the model comes from directly incorporating the social dimension of mobility. Furthermore, as only aggregated call data is required, the model helps to avoid potential privacy problems.

People travel and move for a variety of reasons, including social, economic, and political factors. While individuals may follow simple, recurrent patterns of movement, e.g., daily commuting, a more complex picture emerges when all trajectories of a population are assembled together¹. Understanding the principles governing individual and collective movement is important for a number of reasons: for planning urban design², for forecasting and avoiding traffic congestion³, for mitigating infectious disease^{4–6}, and for contingency planning in extreme situations caused by disasters^{7,8}. However, accurately determining the movement patterns in a population is cumbersome and costly, and involves privacy issues.

There are two ways of inferring the mobility patterns in a population: by direct measurement or by models that predict population movement based on other observed data. Regarding the former, tracking the movement of individuals using location data from mobile phones^{9–11} has emerged as a powerful alternative to traditional methods such as traffic surveys¹². In this case, the data set comes from the billing systems of mobile phone operators, where the closest tower of each phone is recorded when a mobile phone is used. The resolution problems caused by this are compensated by the large quantity and high quality of data^{13,14}. However, there are drawbacks to this approach: tracking the locations of individuals may be seen as a threat to privacy even when the data is properly anonymised¹⁵.

The alternative approach to direct measurement is to use models that predict the average population behaviour from (publicly) available information, such as census and population data. Perhaps the most famous example is the gravity model^{16–18} that has been used to predict the intensity of a number of human interactions, including population movement^{19–21} and mobile phone calls between cities²². In the gravity model, the intensity of interactions between two locations (e.g., cities) is determined by their populations and distance (with proper scaling exponents). Recently, it has been shown that a parameter-free model, the radiation model²³, is able to predict mobility patterns with improved accuracy; this model requires geospatial information on population size as an input.

The applicability of the above-mentioned models is constrained by the availability of accurate population information. This may become a problem e.g. for developing countries, where census data may be incomplete. However, mobile phones are ubiquitous almost everywhere, and one might expect that mobile phone calls reflect the social dimension of mobility – the amount of social ties between geospatial locations can be expected to influence travel patterns. Therefore, the aim of this paper is to predict mobility patterns from mobile phone call data alone, and examine models that would be applicable in a setting where accurate, up-to-date population



information is not available. Furthermore, we focus on models that only require aggregated call data, without needing to track individual users. This has the obvious benefit of mitigating privacy-related issues; additionally, the volume of required input data is smaller and the aggregation can be easily done by the mobile operator that owns the source data.

Our modelling and analysis is purely based on the Ivory Coast mobile telephone data set²⁴, originally released by Orange for the Data for Development Challenge. This data set includes information on mobile phone calls aggregated at the tower level during 140 days, used as inputs for the models, and data on the trajectories of randomly chosen individuals, used for developing the models and testing their accuracy. There is no accurate, up-to-date geospatial population information for Ivory Coast; the last census was conducted in 1998, and there is no data available on mobility or migration within the country. In contrast, the telephone system in Ivory Coast is well-developed by African standards with mobile phone penetration above 83%²⁵.

This paper is constructed as follows: first, we examine gravity laws for average mobility and call frequency between locations. We then proceed to show that mobility between two locations can be directly estimated from the number of calls between the locations and their distance. This holds at two levels of coarse-graining: between tower locations in a major city and between cities. Finally, we study the accuracy of predictions for individual pairs of locations, beyond averages, and show that the number of calls between locations appears to be a good predictor of the frequency of travel between them. For reference, we also study variants of existing mobility models (the gravity and radiation models) where location-specific call frequencies are used as inputs instead of population data; despite applying these models beyond their intended range, they provide fairly good predictions on average.

Results

Data set and coarse-graining. The data set comes in two parts: (i) the number of calls between 1231 Orange towers in Ivory Coast for 5 months, and (ii) ten data sets on two-week individual trajectories of 50,000 randomly chosen users. From the trajectories, we aggregated the mobility m_{ij} between locations i and j by counting direct movements along the trajectories (see Methods for further details).

As it is reasonable to assume that communication and mobility patterns are in general different for short and long distances, we aggregated the data at two levels: (i) tower level for intra-city behaviour and (ii) city level for inter-city behaviour. The intra-city analysis consist of 5.1 million movements and 109 million calls between all 298 towers located inside Abidjan, the largest city of Ivory Coast, during 140 days. This comprises 31% of all calls and 50% of all movements in the country. In this analysis the geographical unit – referred to as “location” in the following – is the area covered by a single tower. To analyse inter-city behaviour, we aggregated towers that lie within a city boundary and consider calls and mobility between cities. The resulting data contains 143 cities with 63 million calls and 374 thousand movements between them during 140 days. At both levels of analysis, we determine the number of calls, movements, and the geographical distance between every pair of locations (towers, cities). See Methods for further details.

Gravity laws: dependence of mobility and communication intensity on distance. We begin by investigating whether the mobility and communication intensities between two locations follow the gravity law on average. In its general form, the gravity law states that

$$x_{ij} \propto \frac{N_i N_j}{d_{ij}^\alpha}, \quad (1)$$

where x_{ij} is the intensity of interaction, e.g., calls, mobility, trade, between locations i and j associated with populations of sizes N_i

and N_j , separated by a distance d_{ij} ^{16–18}. The exponent α governs the distance dependence. Note that in the most general form of the gravity law, N_i and N_j are also associated with an exponent; here for simplicity we assume a linear dependence. For our data, we study the intensities of mobility m_{ij} and communication c_{ij} between locations i and j . These are defined as the average number of weekly movements and calls between them, respectively. As a proxy of the population N_i , we take the total number of weekly calls s_i made and received at location i .

The variation of the scaled mobility intensity, $m_{ij}/s_i s_j$, with respect to the distance d_{ij} is shown in Fig. 1 for the tower and city levels of coarse-graining (panels A and B, respectively). In both cases, the gravity law holds on average and

$$\left\langle \frac{m_{ij}}{s_i s_j} \right\rangle \propto d_{ij}^{-\gamma}, \quad (2)$$

where $\gamma \approx 2.14$ for the intra-city level and $\gamma \approx 2.54$ for the inter-city level. Panels C and D display a similar plot for the scaled communication intensity that is also seen on average to follow the gravity law:

$$\left\langle \frac{c_{ij}}{s_i s_j} \right\rangle \propto d_{ij}^{-\delta}, \quad (3)$$

where the distance exponents are $\delta \approx 1.20$ for the intra-city level and $\delta \approx 1.48$ for the inter-city level. It is worth noting that both exponents γ and δ are smaller for the intra-city level, indicating differences in communication and travel patterns within and between cities: within a city, the spatial distance appears to play a less important role than it does between cities.

The two gravity laws discussed above suggest that the following relationship might also hold:

$$\left\langle \frac{m_{ij}}{c_{ij}} \right\rangle \propto d_{ij}^{-\beta}, \quad (4)$$

where $\beta = \gamma - \delta$. This is indeed the case, as seen in Fig. 1 (E,F) where $\langle m_{ij}/c_{ij} \rangle$ follows a power-law dependence on d_{ij} . For both intra- and inter-city levels, we find the exponent $\beta \approx \gamma - \delta$ (see Table I). These results suggest that there are two possible ways of inferring the intensity of mobility between locations i and j from call data: using the distance and either (i) the total call numbers at both locations s_i and s_j (Eq. 2), or (ii) the total number of calls between the locations c_{ij} (Eq. 4). The prediction accuracy of these two models will be assessed in the section “Prediction accuracy” below.

It is worth noting that both for intra- and inter-city levels, the exponent $\beta \approx 1$. This does not directly result from Eqs. (2) and (3). One possible argument for the observed value of β is as follows: the cost of a single trip, measured in e.g. time or money, between two towers/cities i and j can be assumed to depend linearly on their distance, d_{ij} . This means that the total cost of all movements between i and j is proportional to $m_{ij} d_{ij}$. However, the cost of communication is independent of distance. If one further assumes that the total cost of movement is balanced by the total benefit brought by social ties, linearly reflected in c_{ij} , we have $m_{ij} d_{ij} \sim c_{ij}$ and thus the value of exponent $\beta = 1$. In this interpretation, the communication exponent δ is directly related to a decrease in the number of social ties as function of distance, whereas γ captures a combination of cost associated with travel and the decrease in the number of social ties.

Models for estimating mobility based on call data. The results of the previous section indicate that on average, the mobility intensity m_{ij} between two locations i and j can be estimated using the gravity model

$$m_{ij}^G = k^G \frac{s_i s_j}{d_{ij}^\alpha}, \quad (5)$$

where k^G is a normalization constant obtained by equating the total numbers of expected and observed movements, i.e.,

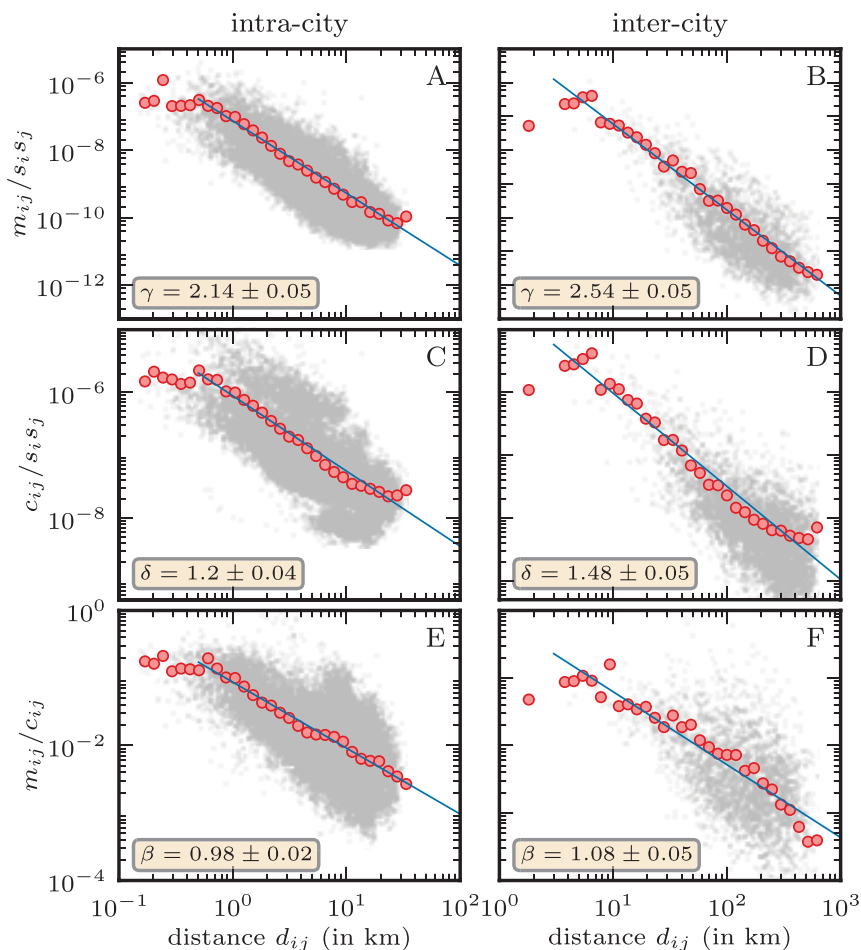


Figure 1 | Dependence of the intensities of interaction on distance. The number of (A,B) movements per strength product $m_{ij}/s_i s_j$, (C,D) calls per strength product $c_{ij}/s_i s_j$ and (E,F) movements per call m_{ij}/c_{ij} decrease with distance between i and j for both intra-city and inter-city analyses. Each grey dot indicates a pair of locations, and circles correspond to the average log-binned behaviour. Solid lines show the fitted power-law decaying behaviour.

$\sum_{ij} m_{ij} = \sum_{ij} m_{ij}^{\text{C}}$. This model takes the communication intensities s_i and s_j at both locations as inputs in addition to the distance d_{ij} . As an alternative we propose the *communication model*

$$m_{ij}^{\text{C}} = k^{\text{C}} \frac{c_{ij}}{d_{ij}^{\beta}}, \quad (6)$$

based on the communication intensity c_{ij} between the locations. The normalization constant k^{C} is obtained as before. The values of the exponents γ and β are taken from Table I.

For comparison, we also study a modified version of the *radiation model*²³, originally designed to predict mobility between locations i and j with the help of data on population density in the surrounding area. Again, we modify the model such that only call and distance data is required as input. To this end, we assume that the number of calls in a given location is an unbiased estimate of population density, similarly to the gravity model. Note that this assumption may not necessarily hold, since mobile phone penetration may correlate with socioeconomic factors. Further, we assume that the number of trips that begin (end) at location i (j) is proportional to s_i (s_j). Then, the radiation model formula can be rewritten as

$$m_{ij}^{\text{R}} = k^{\text{R}} \left[\frac{s_i^2 s_j}{(s_i + s_{ij})(s_i + s_j + s_{ij})} + \frac{s_j s_j^2}{(s_j + s_{ij})(s_j + s_i + s_{ij})} \right]. \quad (7)$$

Here s_{ij} denotes the total number of calls made within a circle of radius d_{ij} centred at i , excluding locations i and j , and k^{R} is a normalization constant.

Prediction accuracy. To assess the actual predictive power of the models beyond averages, we compare the actual mobility intensity m_{ij} , obtained from the trajectory data set, with the estimates given by the models for each specific pair of locations i and j . This comparison for the communication model, the gravity model, and the radiation model is shown in Fig. 2. The gray dots correspond to predicted versus actual mobility for each pair of locations, and the boxes (whiskers) correspond to the region between 25th and 75th (9th and 91st) percentiles.

It is clear from the figure that all models give on average reasonable predictions. However, the gravity and radiation models display higher levels of variance between the predicted and actual mobility intensities. In particular, the prediction accuracy of the gravity model is relatively poor for the inter-city mobility, and the radiation model performs the worst for the intra-city mobility. The latter is not surprising, as the radiation model was originally not designed for predicting short-range travel patterns within cities. Further, the original radiation model requires accurate geospatial population information, and simply equating population size within an area with the number of calls can be expected to give rise to errors.

The level of observed variance implies that in addition to comparing averages, it is important to compare the expected and observed mobility between individual pairs of locations. As the first step, we determine the Spearman correlation coefficients $r^{\text{C,G,R}}$ between m_{ij} and $m_{ij}^{\text{C,G,R}}$. Table II shows that the correlation is higher for the communication model than for the gravity and radiation models for both levels of coarse-graining (intra-city, inter-city). In general,



Table 1 | The estimated values of exponents γ (Eq. 2), δ (Eq. 3), and β (Eq. 4) for the tower and city levels of coarse-graining. The values and their standard errors have been obtained by least square fitting to logarithmically binned data

Level	γ	δ	β
intra-city (tower level)	2.14 ± 0.05	1.20 ± 0.04	0.98 ± 0.02
inter-city (city level)	2.54 ± 0.05	1.48 ± 0.05	1.08 ± 0.05

in terms of the Spearman coefficient, predictions of all models are more accurate for intra-city mobility than for inter-city mobility.

Finally, we consider the differences between the observed and predicted mobilities by measuring their relative deviations. For all the three models, we define the relative deviations $\delta_{ij}^{C,G,R}$ between the observed m_{ij} and predicted $m_{ij}^{C,G,R}$ as

$$\delta_{ij}^{C,G,R} = \frac{m_{ij}^{C,G,R} - m_{ij}}{m_{ij}^{C,G,R} + m_{ij}}, \quad (8)$$

where δ_{ij} takes values between -1 and 1 . A deviation of $\delta_{ij} = 0$ implies exact prediction by the model for the pair of locations i and j , whereas negative (positive) values indicate under- (over-) estimations. We only determine δ_{ij} for those pairs of i and j for which $m_{ij} \neq 0$.

The probability distributions $P(\delta^{C,G,R})$ shown in Fig. 3 confirm the above finding that out of the studied three models for inferring mobility from call data, the communication model has the highest accuracy of prediction. The distribution $P(\delta^C)$ is well centred around zero, whereas especially for inter-city mobility the distributions $P(\delta^G)$ and $P(\delta^R)$ show a bias towards under-estimation. In more detail, for intra-city mobility, the fractions of location pairs with deviations $\delta \in [-0.25, 0.25]$ are 13% for the radiation model, 42% for the gravity model, and 51% for the communication model. For inter-city mobility, the corresponding fractions are 20%, 17% and 33%. Note that for the gravity model, in spite of the fact that the average $\langle m_{ij}/(s_i s_j) \rangle$ follows a $d_{ij}^{-\gamma}$ -dependence (Fig. 1A,B), there is still a significant amount of under-estimation. This indicates that there is a broad distribution of the values of $\langle m_{ij}/(s_i s_j) \rangle$ for a given distance, and the average value is not always a good estimator.

Discussion and conclusion

The goal of this paper has been to investigate simple models that predict the intensities of mobility between two locations on the basis of mobile phone call data and their geospatial distance. The motivation behind this is to provide ways of predicting mobility in situations where accurate information of population size at each location is not available; furthermore, the focus is on aggregated call data, mitigating

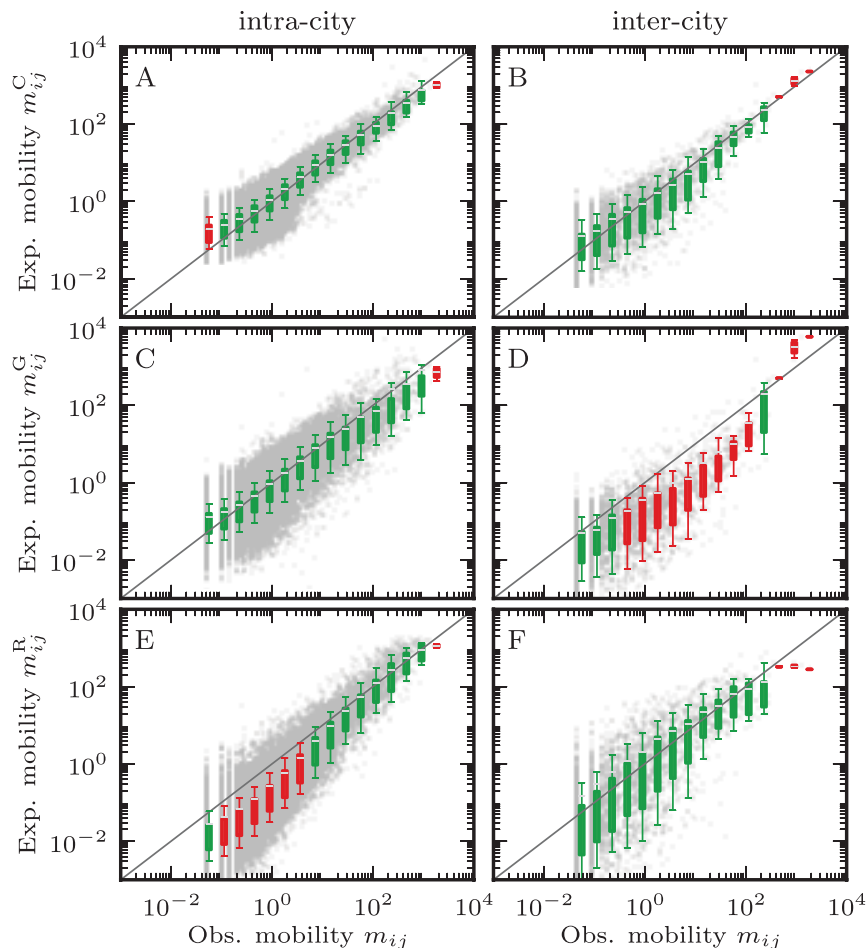


Figure 2 | Comparison between observed and predicted human mobility. The expected mobility intensities (A,B) m_{ij}^C for the communication model, (C,D) m_{ij}^G for the gravity model, and (E,F) m_{ij}^R for the radiation model are plotted against the mobility intensities observed in data m_{ij} . The left panels (A,C,E) correspond to the intra-city analysis and right panels (B,D,F) correspond to inter-city analysis. The boxes provide the region between 25th and 75th percentiles, and the whiskers correspond to 9th and 91st percentiles of logarithmically binned data. A box is colored green if for a given bin the line $y = x$ lies between the 9th and the 91st percentiles of the expected distribution; otherwise it is colored red.



Table II | Spearman correlation coefficient between the observed and predicted mobility values for the three models. For both intra-city and inter-city analyses the communication model shows larger correlation values than gravity and radiation models. The significance of the difference in the correlation is indicated by the p -values

Level	r_s^C	r_s^G	r_s^R	$p(r_s^C > r_s^G)$	$p(r_s^C > r_s^R)$
intra-city (tower level)	0.87	0.81	0.82	$<10^{-4}$	$<10^{-4}$
inter-city (city level)	0.74	0.67	0.67	$<10^{-4}$	$<10^{-4}$

the need to track movement patterns of individual phone users. Our study is based on call and mobility data released by Orange for Ivory Coast; note that it would be important to verify the findings with data from other countries.

We have tested three models that only take aggregated call data and geospatial information as inputs: the well-known gravity model, the communication model based on the number of calls between two locations, and a modified version of the radiation model. While all models on average capture the real mobility patterns derived from call data with location information, a more detailed analysis of the prediction accuracy at the level of individual locations reveals that the communication model is the most accurate out of the three tested models in this setting.

Note that the gravity and radiation models were originally designed to use geospatial population information as input parameters. Since our aim has been to study mobility models in a setting where such information is not available, we have simply taken the number of calls at a given location as a proxy of the population size. Therefore we do not claim that the communication model would outperform other models in a situation where they could be applied as their designers intended. Also note that our modeling target – the mobility pattern – is also derived from mobile phone records, and geospatial biases in mobile phone usage might influence the results. Hence, it would be useful to verify the accuracy of the communication model for a case where there are alternative sources of mobility information.

The likely reason why the communication model works well is that it directly incorporates geospatial information on social ties and human relationships. It has been observed earlier that individuals tend to travel to locations where they have social bonds⁸; furthermore, once under way, it is reasonable to assume that people make calls back home. Because of this, the aggregated intensity of com-

munication between two locations should contain information on the mobility patterns as well. Then, in the first approximation one might assume that the frequency of movement between two locations is directly proportional to the intensity of communication. Further, the simplest way to incorporate the fact that larger distances imply larger travel costs (in terms of time or money) is to assume that mobility is inversely proportional to distance. These two components directly yield the communication model: $m_{ij} \propto c_{ij}/d_{ij}$.

It is worth noting that in general, in gravity laws of human interaction, the distance dependence is associated with some exponent α . This is also seen in our analysis of the gravity laws for mobility and communication intensity, where the exponents were seen to depend on the level of coarse-graining, i.e., intra-city or inter-city. However, for both levels, the inverse distance dependence of the communication model is approximately linear, i.e., the exponent equals one. This suggests universality and calls for analysis of similar data sets from different countries.

Methods

Communication and mobility data. The data set²⁴ consists of 2.5 million call detail records of customers for a single provider (Orange) in Ivory Coast between December 1st, 2011 and April 28th, 2012. The communication data used in this paper contains the number of calls as well as their aggregated duration between all pairs of 1231 towers, i.e., mobile base stations. The geographical locations of the towers were also provided. The temporal resolution of the data set is one hour.

The mobility sample consists of ten data sets of trajectories of individual users, each for 50,000 randomly chosen users. Each trajectory corresponds to the subscribers' call locations during a two-week period. The locations were recorded every time a call was made and correspond to the position of the tower that transmitted the call. The data sets represent consecutive two-week periods, beginning in December 5, 2011.

Determining city boundaries. As the locations of the cell-towers were provided, we used reverse geocoding²⁶ to determine the city in which the tower is located. The mean longitude and latitude of all towers within a city defines the centre of the city. This location was used to calculate the inter-city distances. Out of the 1231 mobile phone towers, 686 are located within city boundaries (with 298 of them in the largest city, Abidjan). The total number of cities with at least a single tower is 143.

Determining direct movements. Given the individual trajectories of users, a variety of methods have been developed to extract different aspects of human mobility¹³. Here, we consider *direct movements* that correspond to any consecutive changes in the location of a user. Formally, direct movements are defined as follows: if the user made a call from location i at some time t and j is the location of the next call at $t' > t$, there is a direct movement from i to j if $j \neq i$. By aggregating this information for all users we determine, the total number of direct movements between all pairs of locations. The locations can correspond either to towers (intra-city analysis) or to cities (inter-city analysis). Note that for inter-city analysis, only towers located within city boundaries are considered. Thus, all calls and direct movements to locations between cities are ignored.

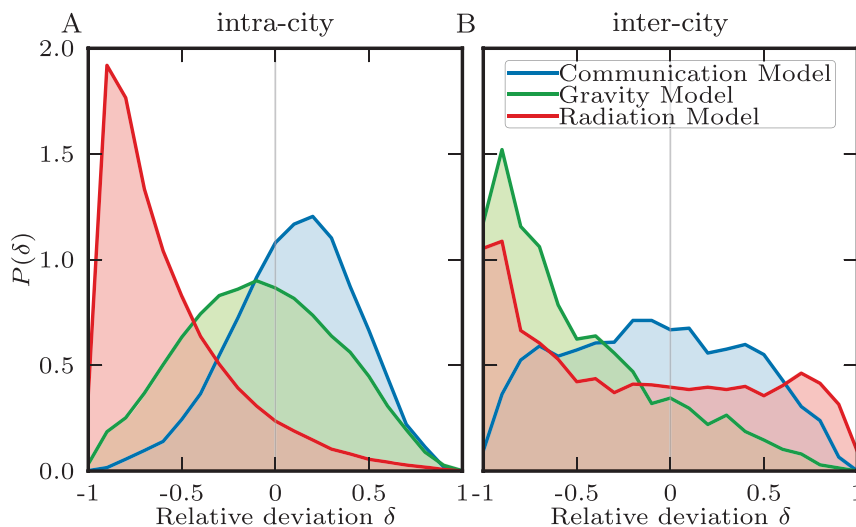


Figure 3 | Relative deviation between the observed and predicted mobility values for the three models. Distribution $P(\delta_{ij}^{C,G,R})$ of the relative deviations $\delta_{ij}^{C,G,R}$ (Eq. 8) for (A) intra-city and (B) inter-city mobility.



Data filtering. Users may be located in areas covered by several towers. In this case, the calls made by users at the same location can be handled by different neighbouring towers. This phenomena of switching of mobile phone calls between towers is called *handover* and it may give rise to artefacts in mobility and communication. For instance, let us consider an *immobile* user located in the boundary area covered by two towers i and j . If one of the calls of this user was served by tower i and the subsequent call by tower j , the data will indicate movement of the user from tower i to tower j . Similarly, the number of calls between neighbouring towers might also get biased. To get rid of this artefact, we excluded all pairs of neighbouring towers from our analysis. As the towers are heterogeneously distributed (higher concentration in densely populated areas and lower concentration in rural zones), neighbouring towers were identified by a distance-independent approach. To do this, we first computed the Voronoi diagram around each tower. The towers having a common edge in their Voronoi cells are defined as the neighbouring towers. We also excluded the communication and mobility between the towers that are located within 1 meter from each other (e.g. two base stations serving a busy area). Further, only pairs of locations with more than one call per day (on average) were considered.

1. Brockmann, D., Hufnagel, L. & Geisel, T. The scaling laws of human travel. *Nature* **439**, 462–465 (2006).
2. Hall, P. *Cities of tomorrow: an intellectual history of urban planning and design in the Twentieth century* (Blackwell, Massachusetts, 2002).
3. Helbing, D. Traffic and related self-driven many-particle systems. *Rev. Mod. Phys.* **73**, 1067–1141 (2001).
4. Balcan, D. *et al.* Multiscale mobility networks and the spatial spreading of infectious diseases. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 21484–21489 (2009).
5. Wesolowski, A. *et al.* Quantifying the impact of human mobility on malaria. *Science* **338**, 267–270 (2012).
6. Dalziel, B. D., Pourbohloul, B. & Ellner, S. P. Human mobility patterns predict divergent epidemic dynamics among cities. *Proc. Natl. Acad. Sci. U.S.A.* **280** (2013).
7. Helbing, D., Farkas, I. & Vicsek, T. Simulating dynamical features of escape panic. *Nature* **407**, 487–490 (2000).
8. Lu, X., Bengtsson, L. & Holme, P. Predictability of population displacement after the 2010 haiti earthquake. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 11576–11581 (2012).
9. Gonzalez, M. C., Hidalgo, C. A. & Barabasi, A.-L. Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008).
10. Song, C., Qu, Z., Blumm, N. & Barabási, A.-L. Limits of predictability in human mobility. *Science* **327**, 1018–1021 (2010).
11. Jo, H.-H., Karsai, M., Karikoski, J. & Kaski, K. Spatiotemporal correlations of handset-based service usages. *EPJ Data Sci.* **1**, 10 (2012).
12. Treiterer, J. Investigation of traffic dynamics by aerial photogrammetry techniques. *Tech. Rep. Ohio Department of Transportation EES 278* (1975).
13. Calabrese, F., Di Lorenzo, G., Liu, L. & Ratti, C. Estimating origin-destination flows using mobile phone location data. *Pervasive Computing, IEEE* **10**, 36–44 (2011).
14. Tizzoni, M. *et al.* On the use of human mobility proxy for the modeling of epidemics. *arXiv, 1309.7272* (2013).
15. Butler, D. Data sharing threatens privacy. *Nature* **449**, 644 (2007).
16. Carey, H. C. *Principles of social science*, vol. 3 (JB Lippincott & Company, 1867).

17. Carrothers, G. A. An historical bedew of the gravity and potential concepts of human interaction. *J. Am. Inst. Plan.* **22**, 94–102 (1956).
18. Anderson, J. E. The gravity model. *Annu. Rev. Econ.* **3**, 133–160 (2011).
19. Barthélemy, M. Spatial networks. *Phys. Rep.* **499**, 1–101 (2010).
20. Jung, W.-S., Wang, F. & Stanley, H. E. Gravity model in the korean highway. *Europhys. Lett.* **81**, 48005 (2008).
21. Thiemann, C., Theis, F., Grady, D., Brune, R. & Brockmann, D. The structure of borders in a small world. *PLoS ONE* **5**, e15422 (2010).
22. Krings, G., Calabrese, F., Ratti, C. & Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *J. Stat. Mech.* L07003 (2009).
23. Simini, F., González, M. C., Maritan, A. & Barabási, A.-L. A universal model for mobility and migration patterns. *Nature* **484**, 96–100 (2012).
24. Blondel, V. D. *et al.* Data for development: the d4d challenge on mobile phone data. *arXiv, 1210.0137* (2012).
25. Cote d ivoire (ivory coast) - telecoms, mobile and broadband - market insights, statistics and forecasts. <http://www.budde.com.au/Research/Cote-d-Ivoire-Ivory-Coast-Telecoms-Mobile-and-Broadband-Market-Insights-Statistics-and-Forecasts.html> (2014) Date of access: 2014-07-10.
26. Reverse geocoding. <https://developers.google.com/maps/documentation/javascript/examples/geocoding-reverse> (2013) Date of access: 2013-01-05.

Acknowledgments

We thank the operator France Telecom-Orange and the “Data for Development” committee for sharing the mobile phone dataset and organizing the D4D challenge. We acknowledge the support by the Academy of Finland, project no. 260427 (JS, RKP) and Aalto University postdoctoral program (HJ). VP was supported by TEKES (FiDiPro). MM was supported in part by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under project no. ON171017. We also acknowledge the computational resources provided by Aalto Science-IT project.

Author contributions

V.P., M.M., H.J., J.S. and R.K.P. designed the research and participated in the writing of the manuscript. V.P., M.M., H.J. and R.K.P. analysed the data and performed the research.

Additional information


Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Palchykov, V., Mitrović, M., Jo, H.-H., Saramäki, J. & Pan, R.K. Inferring human mobility using communication patterns. *Sci. Rep.* **4**, 6174; DOI:10.1038/srep06174 (2014).



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International License. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder in order to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

SCIENTIFIC REPORTS



OPEN

The dynamics of meaningful social interactions and the emergence of collective knowledge

Received: 11 March 2015

Accepted: 11 June 2015

Published: 15 July 2015

Marija Mitrović Dankulov^{1,2}, Roderick Melnik³ & Bosiljka Tadić¹

Collective knowledge as a social value may arise in cooperation among actors whose individual expertise is limited. The process of knowledge creation requires meaningful, logically coordinated interactions, which represents a challenging problem to physics and social dynamics modeling. By combining two-scale dynamics model with empirical data analysis from a well-known Questions & Answers system *Mathematics*, we show that this process occurs as a collective phenomenon in an enlarged network (of actors and their artifacts) where the cognitive recognition interactions are properly encoded. The emergent behavior is quantified by the information divergence and innovation advancing of knowledge over time and the signatures of self-organization and knowledge sharing communities. These measures elucidate the impact of each cognitive element and the individual actor's expertise in the collective dynamics. The results are relevant to stochastic processes involving smart components and to collaborative social endeavors, for instance, crowdsourcing scientific knowledge production with online games.

In modern statistical mechanics¹, it has been recognized that the collective phenomena arise from interactions among the elementary units via a spontaneous transition to an organized state, which can be identified at a larger scale^{2,3}. Recently, this unifying principle is gaining importance in other natural sciences, for instance for elucidating organization in living systems^{4–8}, emergence of coherent activity in neuronal cultures⁹, and developing computational social science¹⁰. In social systems, interactions and cooperations among actors can lead to the recognizable collective behavior, for instance, the development of collective knowledge¹¹, appearance of common norms¹² or language¹³. The quantitative study of the stochastic processes underlying these social phenomena utilizes the methods of statistical physics supported by analysis of the plethora of online empirical data. Some illustrative examples are the appearance of good and bad conduct in online games¹⁴ and groupings induced by the exchange of emotional messages on social sites^{15–18}. However, a deeper understanding of the mechanisms of collaborative social endeavors^{11,19,20} remains a serious challenging problem in physics and social dynamics modeling.

The building of collective knowledge via social interactions is a subtle phenomenon that requires both cognitive elements and an organized effort to solve a particular query. In this stochastic process, the social system that enables transfer of knowledge and the cognitive subsystem are dynamically interlinked and influence each other at a microscopic scale²¹. In the relational epistemology, the exchange of values is an essential factor that permits the emergence of a collective value via interaction and cooperation among equal individuals²². In this concept, the collective knowledge is neither an entity over individuals nor their sum, rather, it is a property of the particular relations among the interacting actors. It reflects the actions of each individual as a *meaningful, adjusted to the actions of others by means of new operation*; its reciprocity and the acceptance of the confirmed values lead to a cooperation “that has a logical structure isomorphic to logical thought”²². On the practical side, modern information communication

¹Department for Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia. ²Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia. ³MS2Discovery Interdisciplinary Research Institute, M2NeT Laboratory and Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada. Correspondence and requests for materials should be addressed to B.T. (email: Bosiljka.Tadic@ijs.si)

technologies (ICT) provide a suitable platform for knowledge building via social dynamics^{23,24}. These systems aim at transferring the expertise and tacit knowledge that reside in the minds of individuals into a form of collective knowledge. Through ICT, the individual's knowledge is shared or “externalized”²¹. Also, the fragile relational state, where the knowledge is dynamically experienced within a community, is actualized as a collection of mutually related digital artifacts. When a systematic tagging is applied to these artifacts, a form of “explicit” knowledge appears, from which others can learn²¹. For this reason, the emergence and quality of the collective knowledge crucially depend on the microscopic mechanisms, by which a particular cognitive element and an individual actor's expertise contribute to the self-organized process.

We develop a new approach that explains how the collective knowledge emerges in Questions & Answers (Q&A) communications. We utilize the concept of two-scale dynamics that enables defining a correct microscopic model of interactions between social and cognitive elements, and confirm its predictions by quantitative analysis of the empirical data from a well-known Q&A system. The elementary units in the process, actors, and questions that they post or answer contain sub-elementary units—cognitive elements, which describe the actor's expertise and the questions' cognitive contents. Their dynamics strictly obeys the cognitive recognition rules, thus influencing the dynamics at the social level of actors. We quantitatively describe the knowledge-creation process from the elementary interactions to mesoscopic and global level. The statistical signatures of the collective dynamics depend on the range of the actors' expertise, which can be extracted from the empirical data and varied in the simulations. The impact of cognitive elements in the empirical data is further confirmed by methods of information theory while the occurrence and structure of communities are visualized by graph theoretic techniques.

Results

Fine-grained dynamics and cooperation. All our exemplifications are provided based on the analysis of data in mathematics from the system known as **Mathematics** which has become a universal clearinghouse for Q&A in the field²⁵. In the data, the cognitive element of each artifact (question, answer or comment) has been systematically tagged according to the standard mathematics classification scheme. In addition, the fact that a unique identity is known for each actor (user) and each artifact together with the high temporal resolution of the data enable a detailed analysis of the underlying stochastic process. Assuming that the cognition-driven events occurred, we determine a set of tags as expertise of each user in the considered dataset. The dataset and the procedure are described in Methods. In the model (Supplementary information, SI), the actors (agents) have a defined range of expertise. Minimal matching of the expertise of an answering agent with the tags of the answered question is strictly obeyed. The considered agents have the activity patterns statistically similar with the patterns of users in the empirical data while their expertise is varied.

In the process, which is schematically depicted in Fig. 1a, an actor (U) posts a question (Q), which may receive answers or comments (A) by other actors over time. Subsequently, new Q and the already present Q&A are subject to further answers, and so on. Representing each action by a directed link, this process co-evolves a bipartite network, where actors are one partition and Q&A form another partition. An example of a single-question network from the empirical data is shown in Fig. 1b. The cognitive content of each question is marked by up to 5 different tags, which thus specify the required expertise of the answering actors. Matching by at least one tag is required. The actor's expertise is transferred to its answer. The excess expertise of the involved actors leads to the innovation^{26–28} and an accumulation of expertise around a particular question. At the same time, it extends the sample space of matching events, thus accelerating the process in a self-organized manner.

The quantitative measures displayed in Fig. 1(c–f) signify a highly cooperative process with the cognitive elements encoded by tags in the empirical dataset. Specifically, the entropy in Fig. 1f shows a distinctly non-random pattern of the appearance of each tag. In accordance with the entropy, the use of different contents shows temporal correlations. The distribution of time intervals between consecutive events with a particular tag ranges over five decades, Fig. 1d, suggesting a variety of roles that different cognitive elements play in the process. The dynamics of tags closely reflects the heterogeneity of the users' activity profile and their expertise. Figure 1d also shows the broad distribution of the interactivity time of a particular user; the presence of a daily cycle is characteristic of online social dynamics^{15,17}. The long delays between actions of some users, contrasted with a frequent activity of others, yield the power-law distribution of the number of activities N_i per user (Fig. 3a in SI). Further, the role of each user in the process can be distinguished. For instance, in Fig. 1c, the probability for posting questions g_i decays with the number of the user's actions N_i . Essential for the cognitive process, however, is the broad range of the user's expertise. As discussed in Methods, it is measured by the entropy distribution shown in Fig. 1e. While the majority expertise includes between one and four tags, few individuals have an activity record for a large number of topics. Consequently, the appearance of a particular combination of cognitive elements shows a complex pattern. All distinct combinations of tags found in the dataset obey Zipf's law, see Fig. 2. It is a marked feature of scale-invariance in the collective dynamics^{28,29}. The ranking distribution of individual tags is also broad, Fig. 2 in SI. Furthermore, by directly inspecting the related time series, Figs 4 and 5, we find that an actively self-organized social process underlies the observed dynamics of cognitive elements.

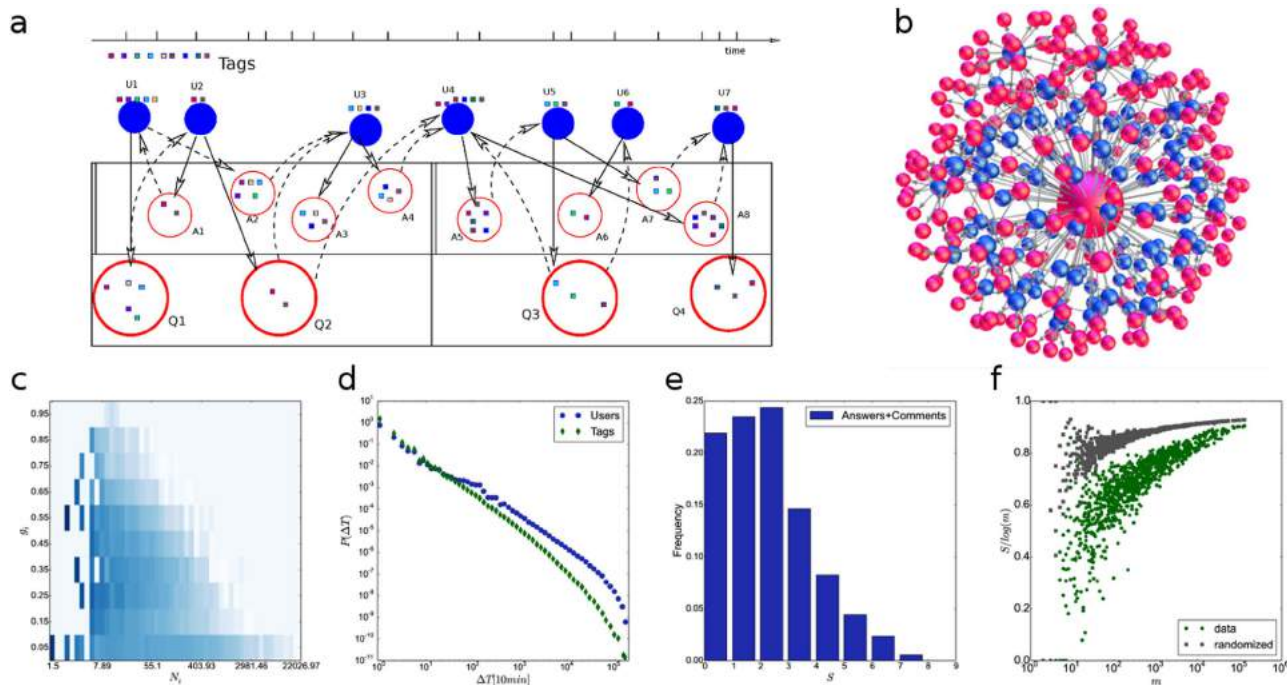


Figure 1. Tags-matching illustration and the activity patterns of users and tags in Mathematics.

(a) Schematically shown a sequence of events with matching of tags (colored boxes) between actors' expertise (displayed as a particular set of tags above blue circles—actors, U_i), the answers A_j , and questions Q_j containing the tags of the related actor's expertise. The direction of lines towards/outwards each actor indicates the process of reading/posting event. (b) Bipartite network of users (blue) and answers (red) at a favorite question (big red node). (c) Probability g_i of posting a new question by the user i plotted against its total activity N_i , averaged over all users in the dataset. (d) The distributions of the interactivity time ΔT for users and tags. (e) The distribution of the user's expertise entropy S_i averaged over all users in the data. (f) Each point indicates the entropy related with the probability of the appearance of a particular tag along a sequence of m time intervals, where m is the tag's frequency. Lower set of points represents the entropies for all tags computed from the sequence of events in the empirical data while the upper set is obtained from its randomized version.

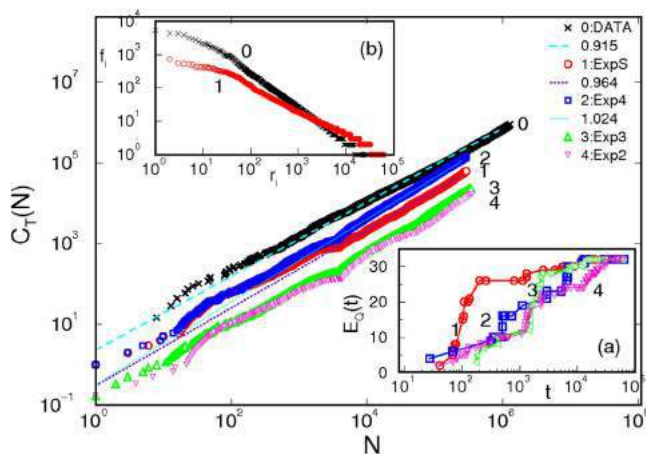


Figure 2. Innovation growth by the actor's expertise. Main panel: The number of new combinations of tags $C_T(N)$ at questions including answers to them is plotted against increasing total number of artifacts N . The curves 0 ... 4 are for the empirical data and simulations where the number of the agent's expertise is fixed as follows: (ExpS), 2^S -tags expertise where S is taken from the distribution in Fig. 1e, and (Exp n), n -tags expertise where $n = 4, 3, 2$. Inset (a) Increase of the knowledge at a particular question $E_Q(t)$ over time t for diverse distributions of expertise as in the central panel. Inset (b) Ranking distribution for frequency of new combinations of tags appearing in questions and the related answers for (0) the empirical data and (1) simulation in the case ExpS.

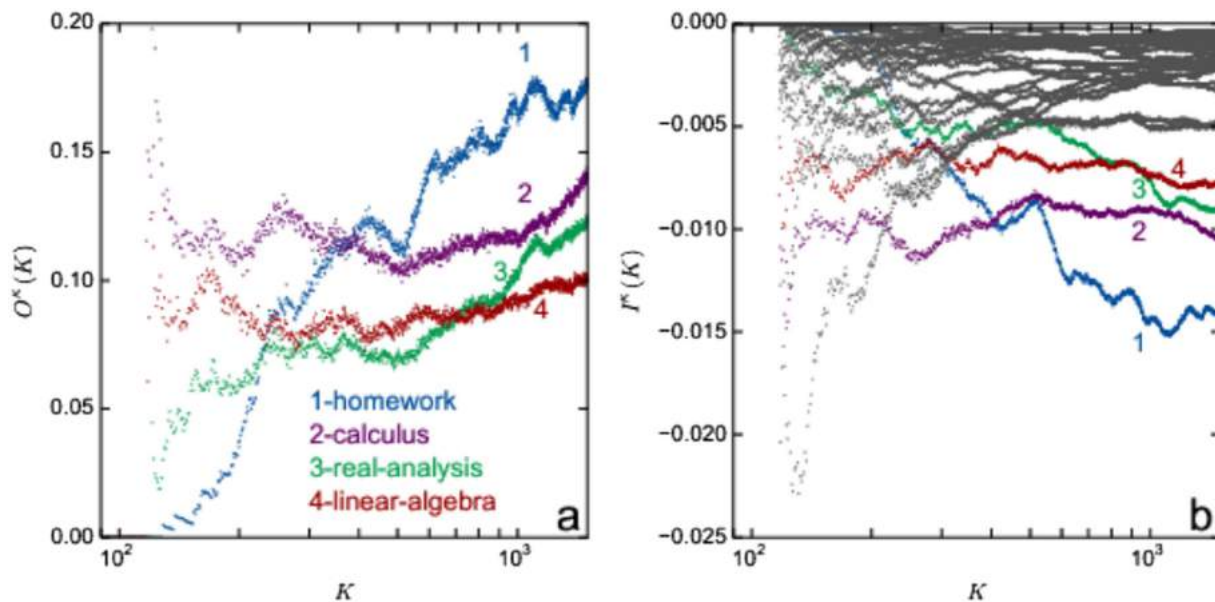


Figure 3. Measuring the impact of a particular cognitive content (κ -tag). Likelihood $O^\kappa(K)$ for four most active tags (a) and Information divergence $I^\kappa(K)$ for 30 most active tags (b) are plotted against the time-window index K .

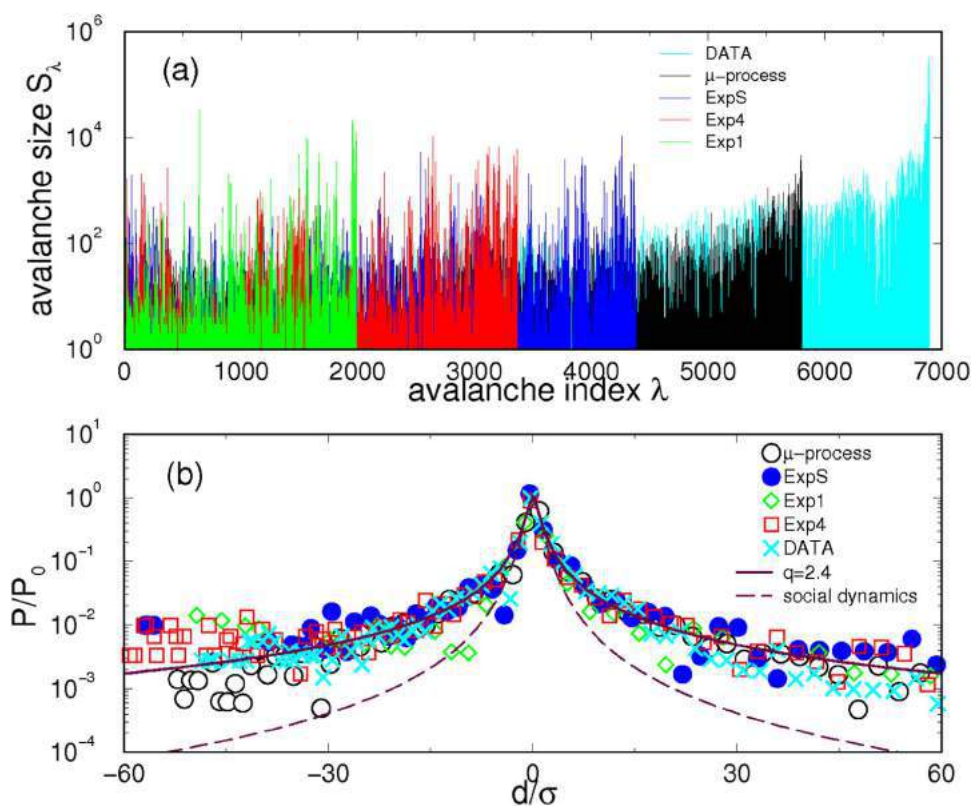


Figure 4. Clustering of events with cognitive contents. (a) Sequences of clustered events (avalanches) of the size S_λ against the cluster's index λ . Different colors indicate the sequences identified in the empirical time series and time series simulated for various ranges of the agents' expertise. (b) Distribution of the first returns scaled by the standard deviation σ of the corresponding sequence (matching color). Full line indicates q -Gaussian curve with the parameter $q=2.4$. For a comparison, the curve with $q=1.8$ is shown (dashed line), corresponding to the case of chat channel dynamics studied in¹⁵.

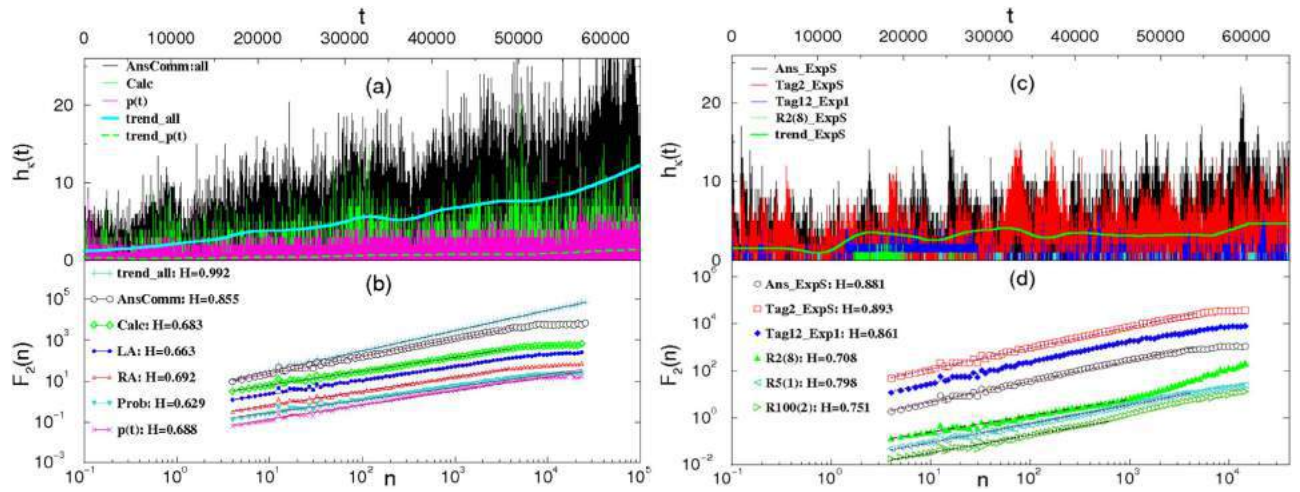


Figure 5. Persistent fluctuations in answering activity in data and in simulations. (a) In the empirical dataset, time series of new users $p(t)$, and time series of the number of answers and comments, and the number of events involving a particular tag (“calculus”). (b) The fluctuations $F_2(n) \sim n^H$ around the local trend are plotted against the time interval n for time series in (a) as well as their trends, and time series involving a particular tag: LA—“linear algebra”, RA—“real algebra”, Prob—“Probability”. (c) and (d) Time series and their fluctuations in the simulations: time series of the number of all answers, and the number of answers containing a particular tag no.2, as well as series containing a particular combination of eight tags R2(8), one-tag, R5(1), and two tags combination, R100(2), all for the distribution of expertise ExpS, and the answers containing tag no.12, in the case of Exp1. Lines are shifted vertically for better display. On each line, the scaling region is indicated by a straight line, whose slope gives the displayed value of the exponent H within error bars ± 0.009 .

Advance of innovation. In the present context, the innovation is measured by appearance of new combination of tags $C_T(N)$ with the addition of artifacts N , Fig. 2. The universal Heaps’ law, $C_T(N) \sim N^\alpha$, and the related²⁹ Zipf’s law shown in the inset (b) of Fig. 2, obtained from the empirical data are supported by the simulations. Here, the innovation is directly given by the excess expertise of the active agents. Thus, the accumulation of expertise at a given question depends on the population of experts; it is slower when, e.g. each agent has two-tags expertise than in the case of four-tags expertise, inset (a) of Fig. 2. In contrast, no increase in innovation is observed when each agent possesses a single-tag expertise.

Information divergence. To examine the influence of a particular cognitive element (tag) in the process, we define a set of conditional probability measures and compute the discrete Kullback-Leibler information divergence from the sequence of question-answer events in which that tag is present, Fig. 3. The empirical data are divided into a series of one-day time windows. In what follows, we use the time window index K , which runs in our examples as $K = 1, 2, \dots, 1498$. As the activity on a particular question or answer typically extends over many time windows, for $K \geq 2$ the space of events Q^K for questions in the K th window also includes Q&A which were active in the $(K-1)$ th window, while only new answers in the K th window make the sample space A^K for answers. By focusing on the time-line of the tag κ , which annotates a particular cognitive content, four conditional probabilities, which are defined in Methods, are determined in every time window K . The information divergence^{30–32}, defined as $I(P(\kappa|A^K, Q^K) || P(\kappa|Q^K)) \equiv I^\kappa(K)$ within the time window K , is computed by

$$I^\kappa(K) = P(\kappa|A^K, Q^K) \ln \frac{P(\kappa|A^K, Q^K)}{P(\kappa|Q^K)}. \tag{1}$$

It determines the information gain about the κ -tag that is present in questions Q^K if the answers A^K are known. Using the chain relation $P(\kappa|A^K, Q^K)P(A^K|Q^K) = P(A^K|\kappa, Q^K)P(\kappa|Q^K)$, it can be expressed as $I^\kappa(K) = P(\kappa|Q^K)O^\kappa(K) \ln O^\kappa(K)$ or:

$$I^\kappa(K) = P(A^K|\kappa, Q^K) \ln \frac{P(A^K|\kappa, Q^K)}{P(A^K|Q^K)} \times \frac{P(\kappa|Q^K)}{P(A^K|Q^K)}, \tag{2}$$

where the ratio $O^\kappa(K) \equiv \frac{P(A^K|\kappa, Q^K)}{P(A^K|Q^K)}$ is a measure of the likelihood that the presence of the κ -tag in questions triggers answers within the time window K . We compute $O^\kappa(K)$ for four most frequent tags,

Fig. 3a, in the sequence of time windows K . A significant difference among tags is apparent; for instance, “real analysis” triggers more activity than “linear algebra”, but still less than “calculus” and “homework” tags.

In view of Eq. (2), the information divergence is expressed (apart from a multiplicative factor smaller than one) as the negative of a relative entropy, which measures the information loss when the probability of answers to questions containing a given cognitive content κ is approximated by the probability of answers to all questions. This probability is expected to increase with the accumulation of expertise around each question over time. Consequently, the information divergence tends to zero for a sufficiently large time. $I^\kappa(K)$, computed for 30 leading tags in the empirical data, Fig. 3b, levels to zero for the majority of tags at large K . However, in the case of four tags, for which the increase in the likelihood of new activity occurs, Fig. 3a, the information divergence still decreases within the entire time interval in the empirical data, four marked curves in Fig. 3b. Note that these topics of a broad interest often combine with new tags, i.e. via the expertise of new arrivals. In this way, triggered answers that match these new tags expand the sample space A^K , which keeps the information divergence finite. This feature is compatible with the innovation growth, reported in Fig. 2. Accounting the contribution of each particular tag in the knowledge creation, the results of information divergence complement the statistical measures in Fig. 1 and support the occurrence of Zipf’s law.

Signatures of self-organization in the social process. The constraints of cognitive recognition at the level of tags affect the social process between actors as well as the structure of the co-evolving network. The time-series analysis is used to uncover prominent features of the coherent fluctuations in this process. We determine the fractal characteristics (see Methods) of the activity time series. In particular, we consider the time series of the number of all answers to the existing questions per time step as well as the time series of such events that contain a particular cognitive element. The results, Fig. 4, reveal that the clustering of events (avalanches) occurs as a distinguishing feature of self-organized processes. In addition, a high persistence is observed in the temporal fluctuations, both in the empirical data and simulations for a varied range of the agent’s expertise, Fig. 5. Measured by Hurst exponent ($H > 0.5$), a similar persistence was found in the processes of thematic discussions¹⁵. While somewhat lower Hurst exponents characterize the fluctuations in prototypical online social interactions¹⁷ and market dynamics³³.

Several sequences of clustered events, determined (see Methods) from the corresponding time series, are reported in Fig. 4a. Considering a particular sequence, the avalanche size differences (returns) $d_\lambda = s_{\lambda+1} - s_\lambda$, $\lambda = 1, 2 \dots \lambda_{max}$ are found to exhibit non-Gaussian fluctuations. Fig. 4b shows the universal plot of the distributions for the appropriately scaled returns. It turns that the q -Gaussian expression $f(x) = a[1 - (1 - q)(x/b)^2]^{1/(1-q)}$, which was observed in a variety of complex dynamical systems^{34–37}, well approximates these distributions (see Fig. 4 in SI). Interestingly, the values³⁶ for the nonextensivity parameter q obtained in these cognitive-driven processes are higher compared with the corresponding parameter in emotion-driven social dynamics¹⁵.

The considered time series and the results of their fractal analysis for the empirical data and simulations are reported in Fig. 5a–d. Note that the rate of new arrivals in the empirical data, $p(t)$, is also used as a creation rate of new agents in the simulation (see Methods). It exhibits distinct temporal correlations, which are carried out from the user’s real life. In this case, $p(t)$ also shows an increasing trend that eventually yields the increase in the entire activity over time both in the empirical and simulated data, Fig. 5a,c. Hence, the detrended fractal analysis is performed, as described in Methods. Shown in Fig. 5b, the fluctuations in the number of answers containing all tags in the empirical data are characterized by the scaling exponent $H = 0.85 \pm 0.07$. Similarly, persistent fluctuations with the exponents in the range $H \in [0.62, 0.68]$ are found in the series of selected events that contain a particular tag. The results of an analogous analysis of the simulated data are shown in Fig. 5d. The time series of the number of answers with all tags and series containing a particular tag have the scaling exponents that are slightly higher, implying a stronger persistence, compared with the corresponding series of the empirical data. Here, we also consider temporal activity of three identified combinations of tags, three bottom curves, which exhibit a similar scaling behavior. These results show that the enhanced self-organization among actors emerges in the interactions with tag recognition, which is mandatory in the model, and, to a large extent, applies to the empirical data.

Knowledge-sharing communities. The coevolving bipartite networks, Fig. 6, emerge in various scenarios in the simulations and empirical data. Note that these networks are different from the single-question graph in Fig. 1b. In this case, each actor is a separate node while a compressed information on a particular question including all answers related to that question represents a single node of the question-partition. The structure of communities detected in these networks clearly stresses the importance of the actor’s expertise. In particular, in the case Exp1, the communities containing a specified single expertise grow as independent clusters, Fig. 6b. The situations when the agents have more than one expertise permit formation of larger communities of agents and questions. For a broad range of the agents’ expertise, the compact communities grow resembling the ones in the empirical data (see also Fig. 5 in SI). It is interesting to note that a dominant node representing a very active knowledgeable

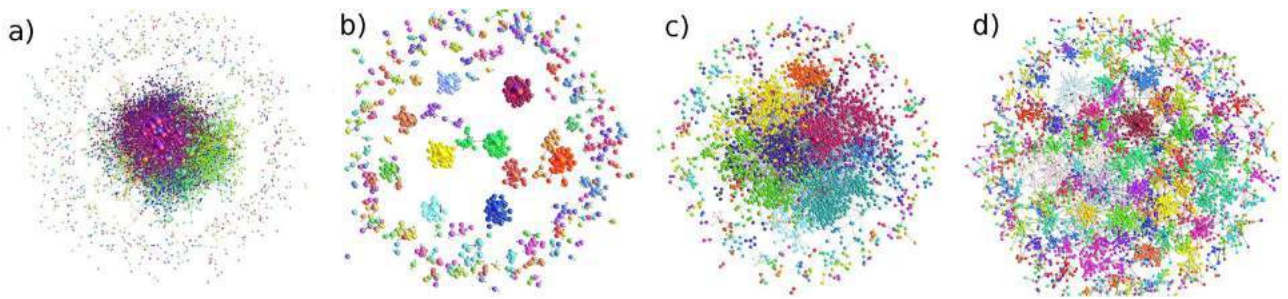


Figure 6. Community structure in bipartite networks of actors and questions reflecting the population of experts. Compressed bipartite network of actors and questions from Mathematics dataset (a) and simulation with the population of experts Exp1 (b) Exp2 (c) and in the case of ExpS but with non matching expertise μ -process (d). The observed communities, indicated by different colors, contain actors interconnected by questions. Each question node contains all answers related to it.

actor appears in each community. On the contrary, the pattern of communities is entirely different when the cognitive recognition does not drive the linking, Fig. 6d.

Conclusions and outlook

Knowledge building via social interactions is studied as a collective phenomenon in an extended space—network of actors and their artifacts, where cognitive recognition interactions are active. We have considered an abundant empirical dataset with cognitive elements as mathematical tags and a two-scale dynamics modeling close to the data, which enabled a quantitative analysis of the process from the microscopic to global scale. Our approach permits to reveal the importance of each cognitive element, as well as the expertise of each actor and its activity pattern in the creation of the collective knowledge. Specifically, when the interacting actors possess a diversity of expertise, the process based on the meaningful (cognitive recognition) interactions leads to the innovation and the advance of knowledge of the emerging communities. When a broad spectrum of expertise is present in the population of actors, i.e. as in the empirical system, the process is quite efficient in creating the enlarged space where innovation can occur. In this case, the formation of coherent communities that share the knowledge is associated with the presence of several actors possessing a broad range of expertise. Notably fewer developed communities and a slower advance of knowledge characterize the population with a narrow distribution of expertise; entirely isolated communities and vanishing of innovation is found in the limiting case of a single expertise per individual. In contrast to the meaningful interactions, the case with ad hoc social linking leads to an entirely different outcome, even though, the individual actors possess a broad distribution of expertise. The advance of innovation measured at the system level appears fragmented in a variety of the emerging communities, each of which shares a limited amount of randomly accumulated knowledge.

The dynamics of social and cognitive elements, interwoven at the elementary scale, induces a type of self-organized process where several quantitative characteristics appear to be different from a prototypical social dynamics. Besides theoretical implications of our results in the study of cognitive-driven processes on networks, the presented approach can be directly applied in the analysis of other empirical systems that entail social collaborative efforts^{19,20}. Examples include, but not limited to, social computing, crowdsourcing scientific knowledge production or scientific discovery games, and other emerging areas of increasing importance in the modern science and society^{38–40}. The presented theoretical concept can prove to be useful in modeling physical systems at nanoscale⁴¹, for instance, the assembly of smart nanostructured materials with biological recognition.

Methods

Data structure. As a platform for scientific collaboration²³, *Mathematics* is a part of *StackExchange: expert answers to your questions* network. For this work, the dataset was downloaded on May 5, 2014 from <https://archive.org/details/stackexchange>. It contains all user-contributed contents on *Mathematics* since the establishment of the site, July 2010, until the end of April 2014. Specifically, the considered dataset contains 77895 users, 269819 questions, 400511 answers and 1265445 comments. A detailed information is given about user id, the user's activity (posting, answering, commenting), time stamp, list of tags for questions, and id of the corresponding question or answer to which a given answer or comment refers. The set of tags in answer/comment is inherited from the related question.

Network mapping and topology analysis. Actions of users in Q&A dynamics are mapped onto a directed bipartite graph, where users, as one partition, interact indirectly via artifacts (questions, answers or comments), as another partition. At a user node i an incoming link is inserted to indicate that that

user reads the corresponding artifact while an outgoing link stands for the user's posting of a new artifact. The path of directed links from a question to a user to answer accurately describes the relationship of the answer to the original question, as it is included in the empirical data and strictly observed in the model. We also introduce a compressed bipartite network, where each question-node includes a question with all answers and comments related to that question; typically, they contain a larger number of tags thus expanding the original question's attributes. The graphs layouts are done using **Gephi**; the community structure is detected by the maximum modularity algorithm⁴².

The user's activity and estimation of expertise. Assuming that a particular expertise of a user i is necessary to answer a given question (which is marked by a set of tags), we consider the amount of the user's actions related to a particular tag, κ . Each tag that appears in the data is considered, in total 1040 tags. Hence, we compute a fraction p_i^κ of the user's actions N_i that is spent at κ -tag. For those tags where p_i^κ exceeds the average probability for that user, we set unity, indicating that the user i is an expert in these categories; thus, the user's i expertise list is formed containing in total n_i^{Exp} tags which received unity mark. The rest of tags receive zeros for that user. The entropy measure for each user, $S_i = -\sum_{\kappa=1}^{N_{tag}} p_i^\kappa \log_2 p_i^\kappa$, remarkably quantifies the heterogeneity of the user's expertise, both in answering and posting questions, Fig. 1e and Fig. 3b in SI, respectively. In the model, the agent's expertise is specified from the list of 32 tags. Different populations of experts correspond to the situations where each agent gets a fixed number n_i^{Exp} tags. In particular, one-tag expertise (Exp1), two-tags expertise (Exp2), four-tags expertise (Exp4), etc., correspond to the agent's expertise list with two, four, etc. randomly selected tags. The case marked as ExpS is close to the empirical data, i.e., each agent gets a list of $2^S \leq 32$ tags, where the random number S is taken from the empirical distribution in Fig. 1e.

Tag-related entropy. Following²⁸, we define T_j as the time interval between the first occurrence of a tag j and the last activity in the dataset. Counting the total number of times m that the tag j occurred, we divide T_j into m equal subintervals. Then for each $i=1, \dots, m$ we count the number of events $f_{ji}(m)$ related to the tag j in the i -th subinterval and compute the entropy $S_j(m)$ of the tag's j sequence as $S_j = -\sum_{i=1}^{i=m} \frac{f_{ji}}{m} \log\left(\frac{f_{ji}}{m}\right)$. For each tag in the dataset, the tag's entropy normalized with the corresponding factor $\log(m)$ is represented by a point in Fig. 1f.

Conditional probabilities of tag-related events. Four conditional probabilities appearing in Eqs (1) and (2), are defined and computed as follows: $P(\kappa|Q^K)$, probability that the κ -tag is present given the presence of questions Q^K , is computed as the frequency of κ -tag in all questions; $P(A^K|\kappa, Q^K)$, the probability that answers A^K exist given the questions Q^K with κ -tag, is given by the fraction of users whose expertise includes κ -tag of all active users in K th window; $P(A^K|Q^K)$, the probability that answers A^K exist given the question Q^K (independently on the presence of κ -tag) is obtained as the ratio of the number of matching tags of all active users in K th window with all tags in the present questions; $P(\kappa|A^K, Q^K)$, the probability to find the tag κ given the questions and answers in the K th window is determined from the above probabilities via chain relation.

Definition of temporally clustered events. A cluster (or avalanche) represents a set of events enclosed between two consecutive drops of the time series to the baseline (noise level)^{43–45}.

Detrended time series analysis. To remove the local trend (an increasing activity and a weak 4-month cycle) appearing in the time series in Fig. 5, we apply the method of overlapping intervals^{17,46}. Then, for each time series $h(k)$, $k=1, 2, \dots, T_h$ the profile $Y(i) = \sum_{k=1}^i (h(k) - \langle h \rangle)$ is divided into N_s segments of length n and the standard deviation around the local trend $y_\mu(i)$ is computed at each segment $\mu=1, 2, \dots, N_s$, i.e., $F_2(\mu, n) = \frac{\sum_{i=1}^n [Y((\mu-1)n+i) - y_\mu(i)]^2}{n}$. Varying the segment length n , the scale invariance $F_2(n) = (1/N_s) \sum_{\mu=1}^{N_s} F_2(\mu, n) \sim n^H$ is examined to determine the Hurst exponent H .

Model rules of interacting agents with expertise. Assuming that the new arrivals in the system boost the activity⁴⁷, the agents are introduced with a pace $p(t)$ agents per time step, where $p(t)$ is the empirical time series of new users, shown in Fig. 5a. Each new agent receives a unique id and a fixed profile. The agents' profiles statistically match the profiles of users in the data. Specifically, the agent's activity level is set by the number of actions $N_i \in P(N_i)$, where $P(N_i)$ is the distribution of the user's activity averaged over all users in the data (see SI:Fig. 3a). Subsequently, the agent's probability g_i to post a question, or otherwise answer other questions, $1 - g_i$, is selected according to the interdependence g_i and N_i shown in Fig. 1c. Furthermore, the agent's expertise is fixed by first setting the number of tags n_i^{Exp} , according to the considered situation, i.e. Exp1, Exp2, Exp4, or ExpS, and then making the list of the agent's expertise of n_i^{Exp} tags by random selection from the common list of 32 tags. The interactivity time of a new agent is set to $\Delta T = 0$, which implies its immediate action. After each completed action, a new delay $\Delta T \in P(\Delta T)$ is taken, where $P(\Delta T)$ is the empirical distribution for users, Fig. 1d. Note that

both $p(t)$ and $P(\Delta T)$ have the same temporal resolution, one bin representing 10 minutes in the original data. All agents are systematically updated, and the agents with an expiring delay time are placed in the *active agents* list. Each active agent, with its probability g_i , puts a new question. Otherwise, it attempts to answer a question from the updated list of interesting questions. The list is created by considering all questions of next-neighbor agents on which an activity occurred within previous $T_0 = 10$ steps. With a given probability that item can be searched elsewhere. In both cases, the agent's action is the subject of the expertise matching. In the case of μ -process, with the probability $\mu = 0.5$ an agent connects to a random question and post an answer while the matching of tags with the agent's expertise is not required, but it can occur by chance (see illustration Fig. 1 in SI, and Algorithm in SI).

References

1. Van Vliet, C. M. *Equilibrium and non-equilibrium statistical mechanics*, 2nd edition, (World Scientific, New Jersey, 2010).
2. Itzykson, C. & Drouffe, J.-M. *Statistical field theory* Volume 1 & 2 (Cambridge University Press, Cambridge, 1989).
3. Balian, R. *From microphysics to macrophysics* Volume I & II (Springer-Verlag, Berlin, 1991 & 1992).
4. Goldenfeld, N. & Woese, C. Biology's next revolution. *Nature*, **445**, 369 (2007).
5. Goldenfeld, N. & Woese, C. Life is physics: evolution as a collective phenomenon far from equilibrium. *Annu. Rev. Condens. Matter Phys.* **2**, 375–399 (2011).
6. Tunstrom, K. *et al.* Collective states, multistability and transitional behavior in schooling fish. *PLoS Comput. Biol.* **9**, e1002915 (2013).
7. Cavagna, A. & Giardina, I. Birds flocks as condensed matter. *Annu. Rev. Condens. Matter Phys.* **5**, Ed. J. S. Langer, 183–207 (2014).
8. Attanasi, A. *et al.* Information transfer and behavioural inertia in starling flocks. *Nature Phys.* **10**, 691–696 (2014).
9. Orlandi, G. J., Soriano, J., Alvarez-Lacalle, E., Teller, S. & Casademunt, J. Noise focusing and the emergence of coherent activity in neuronal cultures. *Nature Phys.* **9**, 582–590 (2013).
10. Conte, R. *et al.*, Manifesto of computational social science, *Eur. Phys. J. Special Topics* **214**, 325–346 (2012).
11. Carpendale, J. I. M. & Müller, U. Editors, *Social Interactions and the Development of Knowledge* (Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey, 2013).
12. Kenrick, D. T., Li, N. P. & Butner, J. Dynamical evolutionary psychology: Individual decision rules and emergent social norms. *Psychol. Rev.* **110**, 3–28 (2003).
13. Loreto, V. & Steels, L. Emergence of language. *Nature Phys.* **3**, 758–760 (2007).
14. Thurner, S., Szell, M. & Sinatra, R. Emergence of good conduct, scaling and Zipf laws in human behavioral sequences in an online world. *PLoS ONE* **7**, e29796 (2012).
15. Tadić, B., Gligorijević, V., Mitrović, M. & Šuvakov, M. Co-evolutionary mechanisms of emotional bursts in online social dynamics and networks. *Entropy* **15**, 5084–5120 (2013).
16. González-Bailón, S., Borge-Holthoefer, J., Rivero, A. & Moreno, Y. The dynamics of protest recruitment through an online network. *Sci. Rep.* **1**, 197 (2011).
17. Šuvakov, M., Mitrović, M., Gligorijević, V. & Tadić, B. How the online social networks are used: dialogues-based structure of myspace. *J. R. Soc. Interface* **10**, 20120819 (2012).
18. von Scheve, Ch. & Salmela, M. Editors *Collective emotions* (Oxford University Press, 2014).
19. Boudreau, K., Gaule, P., Lakhani, K. R., Riedl, Ch. & Woolley, A. From crowd to collaborators: initiating effort and catalyzing interactions among online creative workers. Harvard Business School, Working paper 14-060, (2014), <http://nrs.harvard.edu/urn-3:HUL.InstRepos:12111352>.
20. Lakhani, K. R. & von Hippel, E. How open source software works: “free” user-to-user assistance. *Res. Policy*, **32**, 923–943 (2003).
21. Kimmerle, J., Kress, U. & Held, Ch. The interplay between individual and collective knowledge: technologies for organisational learning. *Knowl. Manag. Res. Pract.* **8**, 33–44 (2010).
22. Kitchener, R. F. Piaget's social epistemology. *Social interactions and the development of knowledge*, J. I. M. Carpendale & U. Müller Editors (Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey, pp. 45–66, 2013).
23. Nielsen, M. *Reinventing discovery: The new era of networked science* (Princeton University Press, 2012).
24. Bowker, G. C., Leigh Star, S., Turner, W. & Gasser, L. Editors, *Social science, technical systems, and cooperative work*, (Psychology Press, New York, 2014).
25. Baez, J. C. Math Blogs. *Notices Amer. Math. Soc.* **333** (2010).
26. Youn, H., Bettencourt, L. M. A., Strumsky, D. & Lobo, J. Invention as a combinatorial problem: evidence from US patents. *arxiv:1406.2938* (2014).
27. Thurner, S., Klimek, P. & Hanel, R. Schumpeterian economic dynamics as a quantifiable minimum model of evolution. *New J. Phys.* **12**, 075029 (2010).
28. Tria, F., Loreto, V., Servedio, V. D. P. & Strogatz, S. H. The dynamics of correlated novelties. *Sci. Rep.*, **4**, 5890 (2014).
29. Font-Clos, F., Boleda, G. & Corral, Á. A scaling law beyond Zipf's law and its relation to Heaps' law. *New J. Phys.* **15**, 093033 (2013).
30. Grendar, M. & Niven, R. K. The Pólya information divergence. *Inform. Science* **180**, 4189–4194 (2010).
31. Niven, R. K. Combinatorial entropies and statistics. *Eur. Phys. J. B* **70**, 49–63 (2009).
32. Hsieh, P.-H. A nonparametric assessment of model adequacy based on Kullback-Leibler divergence. *Stat. Comput.* **23**, 149–162 (2013).
33. Alvarez-Ramirez, J., Alvarez, J., Rodriguez, E. & Fernandez-Anaya, G. Time-varying Hurst exponent for US stock market. *Physica A* **387**, 615–6169 (2008).
34. Tsallis, C. The nonadditive entropy S_q and its applications in physics and elsewhere: Some remarks. *Entropy* **13**, 1765–1804 (2011).
35. Tsallis, C. & Gell-Mann, M. Editors. *Nonextensive entropy—interdisciplinary applications* (Oxford University Press, Oxford, 2004).
36. Pavlos, G. P. *et al.* Universality of Tsallis non-extensive statistics and fractal dynamics of complex systems. *Chaotic Mod. Simul. (CMSIM)* **2**, 395–447 (2012).
37. Caruso, F., Pluchino, A., Latora, V., Vinciguerra, S. & Rapisarda, A. Analysis of self-organized criticality in the Olami-Feder-Christensen model and in real earthquakes. *Phys. Rev. E* **75**, 055101(R) (2007).
38. Fortino, G., Galzarano, S., Gravina, R. & Li, W. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. *Inform. Fusion* **22**, 50–70 (2015).
39. <http://www.gameslearningsociety.org/> Games Learning Society, (2013) Date of access: 14/05/15.
40. <http://www.complex-systems.meduniwien.ac.at/events/insite13/> INSITE Workshop: Games, Science & Society, (2013) Date of access: 14/05/15.
41. Badu, S. R. *et al.* Modeling of RNA nanotubes using molecular dynamics simulations. *Eur. Biophys. J.* **43**, 555–564 (2014).

42. Lancichinetti, A., Kivela, M., Saramäki, J. & Fortunato, S. Characterizing the community structure of complex networks. *PLoS ONE* **5**, e11976 (2010).
43. Tadić, B. Nonuniversal scaling behavior of Barkhausen noise. *Phys. Rev. Lett.* **77**, 3843–3846 (1996).
44. Spasojević, D., Bukvić, S., Milošević, S. & Stanley, H. E. Barkhausen noise: elementary signals, power laws, and scaling relations. *Phys. Rev. E* **54**, 2531 (1996).
45. Mitrović, M., Paltoglou, G. & Tadić, B. Quantitative analysis of bloggers' collective behavior powered by emotions. *J. Stat. Mech. Theor. Exp.* **2011(02)**, P02005 (2011).
46. Hu, J., Gao, J. & Wang, X. Multifractal analysis of sunspot time series: the effects of the 11-year cycle and Fourier truncation. *J. Stat. Mech. Theor. Exp.* **2009(02)**, P02066 (2009).
47. Mitrović, M. & Tadić, B. Dynamics of bloggers' communities: Bipartite networks from empirical data and agent-based modeling *Physica A* **391**, 5264–5278 (2012).

Acknowledgements

This work was supported by the program P1-0044 from the Research Agency of the Republic of Slovenia and from the European Community's COST Action TD1210 KNOWeSCAPE. MMD was supported in part by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under the project no. ON171017. RM thanks for the hospitality during his stay at the Jožef Stefan Institute.

Author Contributions

M.M.D. collected the empirical data, developed the model *in silico* and performed the numerical simulations. B.T. and M.M.D. designed the model *in silico* and performed data analysis. R.M. contributed to the theoretical analysis and the conceptual framework. All authors contributed to data interpretation and wrote the manuscript. B.T. designed and supervised the project.

Additional Information

Supplementary information accompanies this paper at <http://www.nature.com/srep>

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Dankulov, M. M. *et al.* The dynamics of meaningful social interactions and the emergence of collective knowledge. *Sci. Rep.* **5**, 12197; doi: 10.1038/srep12197 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

ARTICLE

Received 19 May 2015 | Accepted 13 Sep 2015 | Published 20 Oct 2015

DOI: 10.1038/ncomms9627

OPEN

Quantifying randomness in real networks

Chiara Orsini^{1,2}, Marija M. Dankulov^{3,4}, Pol Colomer-de-Simón⁵, Almerima Jamakovic⁶, Priya Mahadevan⁷, Amin Vahdat⁸, Kevin E. Bassler^{9,10}, Zoltán Toroczkai¹¹, Marián Boguñá⁵, Guido Caldarelli¹², Santo Fortunato¹³ & Dmitri Krioukov^{1,14}

Represented as graphs, real networks are intricate combinations of order and disorder. Fixing some of the structural properties of network models to their values observed in real networks, many other properties appear as statistical consequences of these fixed observables, plus randomness in other respects. Here we employ the *dk*-series, a complete set of basic characteristics of the network structure, to study the statistical dependencies between different network properties. We consider six real networks—the Internet, US airport network, human protein interactions, technosocial web of trust, English word network, and an fMRI map of the human brain—and find that many important local and global structural properties of these networks are closely reproduced by *dk*-random graphs whose degree distributions, degree correlations and clustering are as in the corresponding real network. We discuss important conceptual, methodological, and practical implications of this evaluation of network randomness, and release software to generate *dk*-random graphs.

¹CAIDA, University of California San Diego, San Diego, California 92093, USA. ²Information Engineering Department, University of Pisa, Pisa 56122, Italy. ³Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Belgrade 11080, Serbia. ⁴Department of Biomedical Engineering and Computational Science, Aalto University School of Science, Helsinki 00076, Finland. ⁵Departament de Física Fonamental, Universitat de Barcelona, Barcelona 08028, Spain. ⁶Communication and Distributed Systems group, Institute of Computer Science and Applied Mathematics, University of Bern, Bern 3012, Switzerland. ⁷Palo Alto Research Center, Palo Alto, California 94304, USA. ⁸Google, Mountain View, California 94043, USA. ⁹Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, Texas 77204, USA. ¹⁰Max Planck Institut für Physik komplexer Systeme, Dresden 01187, Germany. ¹¹Department of Physics and Interdisciplinary Center for Network Science and Applications, University of Notre Dame, Notre Dame, IN 46556, USA. ¹²IMT Alti Studi, Lucca 55100, Italy. ¹³Department of Computer Science, Aalto University School of Science, Helsinki 00076, Finland. ¹⁴Department of Physics, Department of Mathematics, Department of Electrical and Computer Engineering, Northeastern University, Boston, Massachusetts 02115, USA. Correspondence and requests for materials should be addressed to C.O. (email: chiara@caida.org) or to D.K. (email: dima@neu.edu).

Network science studies complex systems by representing them as networks¹. This approach has proven quite fruitful because in many cases the network representation achieves a practically useful balance between simplicity and realism: while always grand simplifications of real systems, networks often encode some crucial information about the system. Represented as a network, the system structure is fully specified by the network adjacency matrix, or the list of connections, perhaps enriched with some additional attributes. This (possibly weighted) matrix is then a starting point of research in network science.

One significant line of this research studies various (statistical) properties of adjacency matrices of real networks. The focus is often on properties that convey useful information about the global network structure that affects the dynamical processes in the system that this network represents². A common belief is that a self-organizing system should evolve to a network structure that makes these dynamical processes, or network functions, efficient^{3–5}. If this is the case, then given a real network, we may ‘reverse engineer’ it by showing that its structure optimizes its function. In that respect the problem of interdependency between different network properties becomes particularly important^{6–10}.

Indeed, suppose that the structure of some real network has property X —some statistically over- or under-represented subgraph, or motif¹¹, for example—that we believe is related to a particular network function. Suppose also that the same network has in addition property Y —some specific degree distribution or clustering, for example—and that all networks that have property Y necessarily have property X as a consequence. Property Y thus enforces or ‘explains’ property X , and attempts to ‘explain’ X by itself, ignoring Y , are misguided. For example, if a network has high density (property Y), such as the interaral cortical network in the primate brain where 66% of edges that could exist do exist¹², then it will necessarily have short path lengths and high clustering, meaning it is a small-world network (property X). However, unlike social networks where the small-world property is an independent feature of the network, in the brain this property is a simple consequence of high density.

The problem of interdependencies among network properties has been long understood^{13,14}. The standard way to address it, is to generate many graphs that have property Y and that are random in all other respects—let us call them Y -random graphs—and then to check if property X is a typical property of these Y -random graphs. In other words, this procedure checks if graphs that are sampled uniformly at random from the set of all graphs that have property Y , also have property X with high probability. For example, if graphs are sampled from the set of graphs with high enough edge density, then all sampled graphs will be small worlds. If this is the case, then X is not an interesting property of the considered network, because the fact that the network has property X is a statistical consequence of that it also has property Y . In this case we should attempt to explain Y rather than X . In case X is not a typical property of Y -random graphs, one cannot really conclude that property X is interesting or important (for some network functions). The only conclusion one can make is that Y cannot explain X , which does not mean however that there is no other property Z from which X follows.

In view of this inherent and unavoidable relativism with respect to a null model, the problem of structure–function relationship requires an answer to the following question in the first place: what is the right base property or properties Y in the null model (Y -random graphs) that we should choose to study the (statistical) significance of a given property X in a given network¹⁵? For most properties X including motifs¹¹, the choice of Y is often just the degree distribution. That is, one usually checks if X is present in random graphs with the same degree

distribution as in the real network. Given that scale-free degree distributions are indeed the striking and important features of many real networks¹, this null model choice seems natural, but there are no rigorous and successful attempts to justify it. The reason is simple: the choice cannot be rigorously justified because there is nothing special about the degree distribution—it is one of infinitely many ways to specify a null model.

Since there exists no unique preferred null model, we have to consider a series of null models satisfying certain requirements. Here we consider a particular realization of such series—the dk -series¹⁶, which provides a complete systematic basis for network structure analysis, bearing some conceptual similarities with a Fourier or Taylor series in mathematical analysis. The dk -series is a converging series of basic interdependent degree- and subgraph-based properties that characterize the local network structure at an increasing level of detail, and define a corresponding series of null models or random graph ensembles. These random graphs have the same distribution of differently sized subgraphs as in a given real network. Importantly, the nodes in these subgraphs are labelled by node degrees in the real network. Therefore, this random graph series is a natural generalization of random graphs with fixed average degree, degree distribution, degree correlations, clustering and so on. Using dk -series we analyse six real networks, and find that they are essentially random as soon as we constrain their degree distributions, correlations, and clustering to the values observed in the real network ($Y = \text{degrees} + \text{correlations} + \text{clustering}$). In other words, these basic local structural characteristics almost fully define not only local but also global organization of the considered networks. These findings have important implications on research dealing with network structure–function interplay in different disciplines where networks are used to represent complex natural or designed systems. We also find that some properties of some networks cannot be explained by just degrees, correlations, and clustering. The dk -series methodology thus allows one to detect which particular property in which particular network is non-trivial, cannot be reduced to basic local degree- or subgraph-based characteristics, and may thus be potentially related to some network function.

Results

General requirements to a systematic series of properties. The introductory remarks above instruct one to look not for a single base property Y , which cannot be unique or universal, but for a systematic series of base properties Y_0, Y_1, \dots . By ‘systematic’ we mean the following conditions: (1) inclusiveness, that is, the properties in the series should provide strictly more detailed information about the network structure, which is equivalent to requiring that networks that have property Y_d (Y_d -random graphs), $d > 0$, should also have properties $Y_{d'}$ for all $d' = 0, 1, \dots, d - 1$; and (2) convergence, that is, there should exist property Y_D in the series that fully characterizes the adjacency matrix of any given network, which is equivalent to requiring that Y_D -random graphs is only one graph—the given network itself. If these Y -series satisfy the conditions above, then whatever property X is deemed important now or later in whatever real network, we can always standardize the problem of explanation of X by reformulating it as the following question: what is the minimal value of d in the above Y -series such that property Y_d explains X ? By convergence, such d should exist; and by inclusiveness, networks that have property $Y_{d'}$ with any $d' = d, d + 1, \dots, D$, also have property X . Assuming that properties Y_d are once explained, the described procedure provides an explanation of any other property of interest X .

The general philosophy outlined above is applicable to undirected and directed networks, and it is shared by different

approaches, including motifs¹¹, graphlets¹⁷ and similar constructions¹⁸, albeit they violate the inclusiveness condition as we show below. Yet one can still define many different Y -series satisfying both conditions above. Some further criteria are needed to focus on a particular one. One approach is to use degree-based tailored random graphs as null models for both undirected^{19–21} and directed^{22,23} networks. The criteria that we use to select a particular Y -series in this study are simplicity and the importance of subgraph- and degree-based statistics in networks. Indeed, in the network representation of a system, subgraphs, their frequency and convergence are the most natural and basic building blocks of the system, among other things forming the basis of the rigorous theory of graph family limits known as graphons²⁴, while the degree is the most natural and basic property of individual nodes in the network. Combining the subgraph- and degree-based characteristics leads to dk -series¹⁶.

dk -series. In dk -series, properties Y_d are dk -distributions. For any given network G of size N , its dk -distribution is defined as a collection of distributions of G 's subgraphs of size $d = 0, 1, \dots, N$ in which nodes are labelled by their degrees in G . That is, two isomorphic subgraphs of G involving nodes of different degrees—for instance, edges ($d = 2$) connecting nodes of degrees 1, 2 and 2, 2—are counted separately. The $0k$ -distribution is defined as the average degree of G . Figure 1 illustrates the dk -distributions of a graph of size 4.

Thus defined the dk -series subsumes all the basic degree-based characteristics of networks of increasing detail. The zeroth element in the series, the $0k$ -distribution, is the coarsest characteristic, the average degree. The next element, the $1k$ -distribution, is the standard degree distribution, which is the number of nodes—subgraphs of size 1—of degree k in the network. The second element, the $2k$ -distribution, is the joint degree distribution, the number of subgraphs of size 2—edges—between nodes of degrees k_1 and k_2 . The $2k$ -distribution thus defines 2-node degree correlations and network's assortativity. For $d = 3$, the two non-isomorphic subgraphs are triangles and wedges, composed of nodes of degrees k_1, k_2 and k_3 , which defines clustering, and so on. For arbitrary d the dk -distribution characterizes the d -degree k -relations in d -sized subgraphs, thus including, on the one hand, the correlations of degrees of nodes located at hop distances below d , and, on the other hand, the statistics of d -cliques in G . We will also consider dk -distributions with fractional $d \in (2, 3)$ which in addition to specifying two-node degree correlations ($d = 2$), fix some $d = 3$ substatistics related to clustering.

The dk -series is inclusive because the $(d + 1)k$ -distribution contains the same information about the network as the dk -distribution, plus some additional information. In the simplest $d = 0$ case for example, the degree distribution $P(k)$ ($1k$ -distribution) defines the average \bar{k} ($0k$ -distribution) via $\bar{k} = \sum_k kP(k)$. The analogous expression for $d = 1, 2$ are derived in Supplementary Note 1.

It is important to note that if we omit the degree information, and just count the number of d -sized subgraphs in a given network regardless their node degrees, as in motifs¹¹, graphlets¹⁷ or similar constructions¹⁸, then such degree- k -agnostic d -series (versus dk -series) would not be inclusive (Supplementary Discussion). Therefore, preserving the node degree (k) information is necessary to make a subgraph-based (d) series inclusive. The dk -series is clearly convergent because at $d = N$, where N is the network size, the Nk -distribution fully specifies the network adjacency matrix.

A sequence of dk -distributions then defines a sequence of random graph ensembles (null models). The dk -graphs are a set

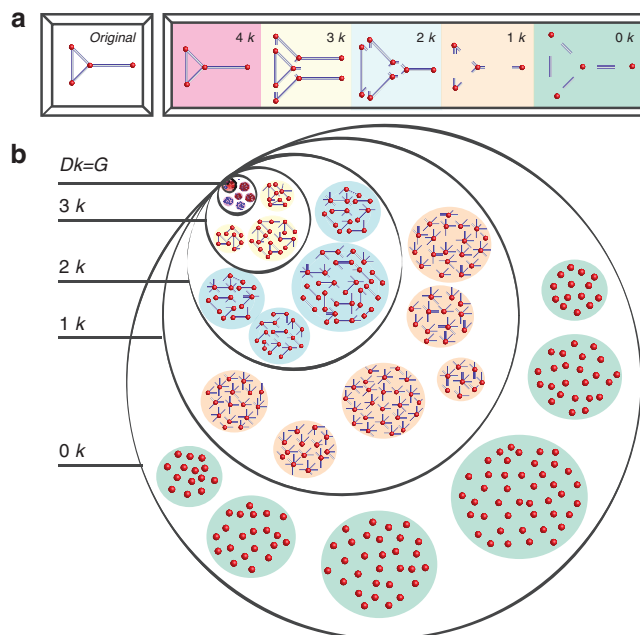


Figure 1 | The dk -series illustrated. (a) shows the dk -distributions for a graph of size 4. The $4k$ -distribution is the graph itself. The $3k$ -distribution consists of its three subgraphs of size 3: one triangle connecting nodes of degrees 2, 2 and 3, and two wedges connecting nodes of degrees 2, 3 and 1. The $2k$ -distribution is the joint degree distribution in the graph. It specifies the number of links (subgraphs of size 2) connecting nodes of different degrees: one link connects nodes of degrees 2 and 2, two links connect nodes of degrees 2 and 3, and one link connects nodes of degree 3 and 1. The $1k$ -distribution is the degree distribution in the graph. It lists the number of nodes (subgraphs of size 1) of different degree: one node of degree 1, two nodes of degree 2, and one node of degree 3. The $0k$ -distribution is just the average degree in the graph, which is 2. (b) illustrates the inclusiveness and convergence of dk -series by showing the hierarchy of dk -graphs, which are graphs that have the same dk -distribution as a given graph G of size D . The black circles schematically show the sets of dk -graphs. The set of $0k$ -graphs, that is, graphs that have the same average degree as G , is largest. Graphs in this set may have a structure drastically different from G 's. The set of $1k$ -graphs is a subset of $0k$ -graphs, because each graph with the same degree distribution as in G has also the same average degree as G , but not vice versa. As a consequence, typical $1k$ -graphs, that is, $1k$ -random graphs, are more similar to G than $0k$ -graphs. The set of $2k$ -graphs is a subset of $1k$ -graphs, also containing G . As d increases, the circles become smaller because the number of different dk -graphs decreases. Since all the dk -graph sets contain G , the circles 'zoom-in' on it, and while their number decreases, dk -graphs become increasingly more similar to G . In the $d = D$ limit, the set of Dk -graphs consists of only one element, G itself.

of all graphs with a given dk -distribution, for example, with the dk -distribution in a given real network. The dk -random graphs are a maximum-entropy ensemble of these graphs¹⁶. This ensemble consists of all dk -graphs, and the probability measure is uniform (unbiased): each graph G in the ensemble is assigned the same probability $P(G) = 1/\mathcal{N}_d$, where \mathcal{N}_d is the number of dk -graphs. For $d = 0, 1, 2$ these are well studied classical random graphs $\mathcal{G}_{N,M}$ (ref. 25), configuration model^{26–28} and random graphs with a given joint degree distribution²⁹, respectively. Since a sequence of dk -distributions is increasingly more informative and thus constraining, the corresponding sequence of the sizes of dk -random graph ensembles is non-increasing and shrinking to 1, $\mathcal{N}_0 \geq \mathcal{N}_1 \geq \dots \geq \mathcal{N}_N = 1$, Fig. 1. At low $d = 0, 1, 2$ these numbers \mathcal{N} can be calculated either exactly or approximately^{30,31}.

We emphasize that in dk -graphs the dk -distribution constraints are sharp, that is, they hold exactly—all dk -graphs have exactly the same dk -distribution. An alternative description uses soft maximum-entropy ensembles belonging to the general class of exponential random graph models^{32–35} in which these constraints hold only on average over the ensemble—the expected dk -distribution in the ensemble (not in any individual graph) is fixed to a given distribution. This ensemble consists of all possible graphs G of size N , and the probability measure $P(G)$ is the one maximizing the ensemble entropy $S = -\sum_G P(G) \ln P(G)$ under the dk -distribution constraints. Using analogy with statistical mechanics, sharp and soft ensemble are often called microcanonical and canonical, respectively.

As a consequence of the convergence and inclusiveness properties of dk -series, any network property X of any given network G is guaranteed to be reproduced with any desired accuracy by high enough d . At $d=N$ all possible properties are reproduced exactly, but the Nk -graph ensemble trivially consists of only one graph, G_{self} , and has zero entropy. In the sense that the entropy of dk -ensembles $S_d = \ln \mathcal{N}_d$ is a non-increasing function of d , the smaller the d , the more random the dk -random graphs, which also agrees with the intuition that dk -random graphs are ‘the less random and the more structured’, the higher the d . Therefore, the general problem of explaining a given property X reduces to the general problem of how random a graph ensemble must be so that X is statistically significant. In the dk -series context, this question becomes: how much local degree information, that is, information about concentrations of degree-labelled subgraphs of what minimal size d , is needed to reproduce a possibly global property X with a desired accuracy?

Below we answer this question for a set of popular and commonly used structural properties of some paradigmatic real networks. But to answer this question for any property in any network, we have to be able to sample graphs uniformly at random from the sets of dk -graphs—the problem that we discuss next.

dk -random graph sampling. Soft dk -ensembles tend to be more amenable for analytic treatment, compared with sharp ensembles, but even in soft ensembles the exact analytic expressions for expected values are known only for simplest network properties in simplest ensembles^{36,37}. Therefore, we retreat to numeric experiments here. Given a real network G , there exist two ways to sample dk -random graphs in such experiments: dk -randomize G generalizing the randomization algorithms in refs 38,39, or construct random graphs with G 's dk -sequence from scratch^{16,40}, also called direct construction^{41–44}.

The first option, dk -randomization, is easier. It accounts for swapping random (pairs of) edges, starting from G , such that the dk -distribution is preserved at each swap, Fig. 2. There are many concerns with this prescription⁴⁵, two of which are particularly important. The first concern is if this process ‘ergodic’, meaning that if any two dk -graphs are connected by a chain of dk -swaps. For $d=1$ the two-edge swap is ergodic^{38,39}, while for $d=2$ it is not ergodic. However, the so-called restricted two-edge swap, when at least one node attached to each edge has the same degree, Fig. 2, was proven to be ergodic⁴⁶. It is now commonly believed that there is no edge-swapping operation, of this or other type, that is ergodic for the $3k$ -distribution, although a definite proof is lacking at the moment. If there exists no ergodic $3k$ -swapping, then we cannot really rely on it in sampling dk -random graphs because our real network G can be trapped on a small island of atypical dk -graphs, which is not connected by any dk -swap chain to the main land of many typical dk -graphs. Yet we note that in

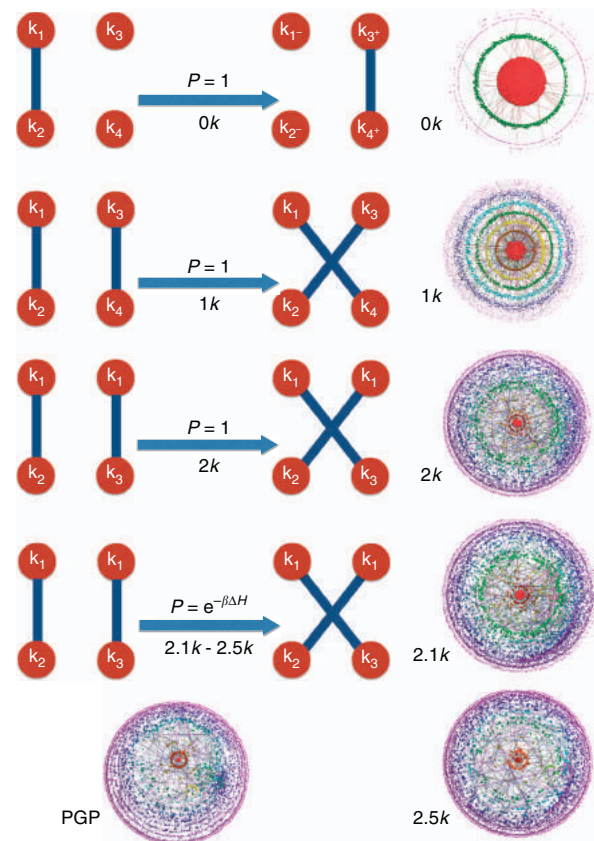


Figure 2 | The dk -sampling and convergence of dk -series illustrated. The left column shows the elementary swaps of dk -randomizing (for $d=0, 1, 2$) and dk -targeting (for $d=2.1, 2.5$) rewiring. The nodes are labelled by their degrees, and the arrows are labelled by the rewiring acceptance probability. In dk -randomizing rewiring, random (pairs of) edges are rewired preserving the graph's dk -distribution (and consequently its $d'k$ -distributions for all $d' < d$). In $2.1k$ - and $2.5k$ -targeting rewiring, the moves preserve the $2k$ -distribution, but each move is accepted with probability p designed to drive the graph closer to a target value of average clustering \bar{c} ($2.1k$) or degree-dependent clustering $\bar{c}(k)$ ($2.5k$): $p = \min(1, e^{-\beta\Delta H})$, where β the inverse temperature of this simulated annealing process, $\Delta H = H_a - H_b$, and $H_{a,b}$ are the distances, after and before the move, between the current and target values of clustering: $H_{2.1k} = |\bar{c}_{\text{current}} - \bar{c}_{\text{target}}|$ and $H_{2.5k} = \sum_i |\bar{c}_{\text{current}}[k_i] - \bar{c}_{\text{target}}[k_i]|$. The right column shows LaNet-vi (ref. 65) visualizations of the results of these dk -rewiring processes (Supplementary Methods), applied to the PGP network, visualized at the bottom of the left column. The node sizes are proportional to the logarithm of their degrees, while the colour reflects node coarseness⁶⁵. As d grows, the shown dk -random graphs quickly become more similar to the real PGP network.

an unpublished work⁴⁷ we showed that five out of six considered real networks were virtually indistinguishable from their $3k$ -randomizations across all the considered network properties, although one network (power grid) was very different from its $3k$ -random counterparts.

The second concern with dk -randomization is about how close to uniform sampling the dk -swap Markov chain is after its mixing time is reached—its mixing time is yet another concern that we do not discuss here, but according to many numerical experiments and some analytic estimates, it is $O(M)$ ^{16,29,38–40,46}. Even for $d=1$ the swap chain does not sample $1k$ -graphs uniformly at random, yet if the edge-swap process is done correctly, then the sampling is uniform^{20,21}.

A simple algorithm for the second dk -sampling option, constructing dk -graphs from scratch, is widely known for $d=1$: given G 's degree sequence $\{k_i\}$, build a $1k$ -random graph by attaching k_i half-edges ('stubs') to node i , and then connect random pairs of stubs to form edges²⁷. If during this process a self-loop (both stubs are connected to the same node) or double-edge (two edges between the same pair of nodes) is formed, one has to restart the process from scratch since otherwise the graph sampling is not uniform⁴⁸. If the degree sequence is power-law distributed with exponent close to -2 as in many real networks, then the probability that the process must be restarted approaches 1 for large graphs⁴⁹, so that this construction process never succeeds. An alternative greedy algorithm is described in ref. 42, which always quickly succeeds and gives an efficient way of testing whether a given sequence of integers is graphical, that is, whether it can be realized as a degree sequence of a graph. The base sampling procedure does not sample graphs uniformly, but then an importance sampling procedure is used to account for the bias, which results in uniform sampling. Yet again, if the degree distribution is a power law, then one can show that even without importance sampling, the base sampling procedure is uniform, since the distribution of sampling weights that one can compute for this greedy algorithm approaches a delta function. Extensions of the naive $1k$ -construction above to $2k$ are less known, but they exist^{16,29,44,50}. Most of these $2k$ -constructions do not sample $2k$ -graphs exactly uniformly either⁴⁶, but importance sampling^{20,44} can correct for the sampling biases.

Unfortunately, to the best of our knowledge, there currently exists no $3k$ -construction algorithm that can be successfully used in practice to generate large $3k$ -graphs with $3k$ -distributions of real networks. The $3k$ -distribution is quite constraining and non-local, so that the dk -construction methods described above for $d=1, 2$ cannot be readily extended to $d=3$ (ref. 16). There is yet another option— $3k$ -targeting rewiring, Fig. 2. It is $2k$ -preserving rewiring in which each $2k$ -swap is accepted not with probability 1, but with probability equal to $\min(1, \exp(-\beta\Delta H))$, where β is the inverse temperature of this simulated-annealing-like process, and ΔH is the change in the L^1 distance between the $3k$ -distribution in the current graph and the target $3k$ -distribution before and after the swap. This probability favours and, respectively, suppresses $2k$ -swaps that move the graph closer or farther from the target $3k$ -distribution. Unfortunately, we report that in agreement with⁴⁰ this $2k$ -preserving $3k$ -targeting process never converged for any considered real network—regardless of how long we let the rewiring code run, after the initial rapid decrease, the $3k$ -distance, while continuing to slowly decrease, remained substantially large. The reason why this process never converges is that the $3k$ -distribution is extremely constraining, so that the number of $3k$ -graphs \mathcal{N}_3 is infinitesimally small compared with the number of $2k$ -graphs \mathcal{N}_2 , $\mathcal{N}_3/\mathcal{N}_2 \ll 1$ (refs 16,30). Therefore, it is extremely difficult for the $3k$ -targeting Markov chain to find a rare path to the target $3k$ -distribution, and the process gets hopelessly trapped in abundant local minima in distance H .

Therefore, on one hand, even though $3k$ -randomized versions of many real networks are indistinguishable from the original networks across many metrics⁴⁷, we cannot use this fact to claim that at $d=3$ these metrics are not statistically significant in those networks, because the $3k$ -randomization Markov chain may be non-ergodic. On the other hand, we cannot generate the corresponding $3k$ -random graphs from scratch in a feasible amount of compute time. The $3k$ -random graph ensemble is not analytically tractable either. Given that $d=2$ is not enough to guarantee the statistical insignificance of some important properties of some real networks, see ref. 47 and below, we, as in ref. 40, retreat to numeric investigations of $2k$ -random graphs

in which in addition to the $2k$ -distribution, some substatistics of the $3k$ -distribution is fixed. Since strong clustering is a ubiquitous feature of many real networks¹, one of the most interesting such substatistics is clustering.

Specifically we study $2.1k$ -random graphs, defined as $2k$ -random graphs with a given value of average clustering \bar{c} , and $2.5k$ -random graphs, defined as $2k$ -random graphs with given values of average clustering $\bar{c}(k)$ of nodes of degree k (ref. 40). The $3k$ -distribution fully defines both $2.1k$ - and $2.5k$ -statistics, while $2.5k$ defines $2.1k$. Therefore, $2k$ -graphs are a superset of $2.1k$ -graphs, which are a superset of $2.5k$ -graphs, which in turn contain all the $3k$ -graphs, $\mathcal{N}_2 > \mathcal{N}_{2.1} > \mathcal{N}_{2.5} > \mathcal{N}_3$. Therefore if a particular property is not statistically significant in $2.5k$ -random graphs, for example, then it is not statistically significant in $3k$ -random graphs either, while the converse is not generally true.

We thus generate 20 dk -random graphs with $d=0, 1, 2, 2.1, 2.5$ for each considered real network. For $d=0,1,2$ we use the standard dk -randomizing swapping, Fig. 2. We do not use its modifications to guarantee exactly uniform sampling^{20,21}, because: (1) even without these modifications the swapping is close to uniform in power-law graphs, (2) these modifications are non-trivial to efficiently implement, and (3) we could not extend these modifications to the $2.1k$ and $2.5k$ cases. As a consequence, our sampling is not exactly uniform, but we believe it is close to uniform for the reasons discussed above. To generate dk -random graphs with $d=2.1, 2.5$, we start with a $2k$ -random graph, and apply to it the standard $2k$ -preserving $2.xk$ -targeting ($x=1, 5$) rewiring process, Fig. 2. The algorithms that do that, as described in ref. 40, did not converge on some networks, so that we modified the algorithm in ref. 10 to ensure the convergence in all cases. The details of these modifications are in Supplementary Methods (the parameters used are listed in Supplementary Table 4), along with the details of the software package implementing these algorithms that we release to public⁵¹.

Real versus dk -random networks. We performed an extensive set of numeric experiments with six real networks—the US air transportation network, an fMRI map of the human brain, the Internet at the level of autonomous systems, a technosocial web of trust among users of the distributed Pretty Good Privacy (PGP) cryptosystem, a human protein interaction map, and an English word adjacency network (Supplementary Note 2 and Supplementary Table 3 present the analysed networks). For each network we compute its average degree, degree distribution, degree correlations, average clustering, averaging clustering of nodes of degree k and based on these dk -statistics generate a number of dk -random graphs as described above for each $d=0, 1, 2, 2.1, 2.5$. Then for each sample we compute a variety of network properties, and report their means and deviations for each combination of the real network, d , and the property. Figures 3–6 present the results for the PGP network; Supplementary Note 3, Supplementary Figs 1–10, and Supplementary Tables 1–2 provide the complete set of results for all the considered real networks. The reason why we choose the PGP network as our main example is that this network appears to be 'least random' among the considered real networks, in the sense that the PGP network requires higher values of d to reproduce its considered properties. The only exception is the brain network. Some of its properties are not reproduced even by $d=2.5$.

Figure 2 visualizes the PGP network and its dk -randomizations. The figure illustrates the convergence of dk -series applied to this network. While the $0k$ -random graph has very little in common with the real network, the $1k$ -random one is somewhat more similar, even more so for $2k$, and there is very little visual

difference between the real PGP network and its 2.5k-random counterpart. This figure is only an illustration though, and to have a better understanding of how similar the network is to its randomization, we compare their properties.

We split the properties that we compare into the following categories. The microscopic properties are local properties of individual nodes and subgraphs of small size. These properties can be further subdivided into those that are defined by the dk -distributions—the degree distribution, average neighbour degree, clustering, Fig. 3—and those that are not fixed by the dk -distributions—the concentrations of subgraphs of size 3 and 4, Fig. 4. The mesoscopic properties— k -coreness and k -density (the latter is also known as m -coreness or edge multiplicity, Supplementary Note 1), Fig. 5—depend both on local and global aspects of network organization. Finally, the macroscopic properties are truly global ones—betweenness, the distribution of hop lengths of shortest paths, and spectral properties, Fig. 6. In Supplementary Note 3 we also report some extremal properties, such as the graph diameter (the length of the longest shortest

path), and Kolmogorov–Smirnov distances between the distributions of all the considered properties in real networks and their corresponding dk -random graphs. The detailed definitions of all the properties that we consider can be found in Supplementary Note 1.

In most cases—henceforth by ‘case’ we mean a combination of a real network and one of its considered property—we observe a nice convergence of properties as d increases. In many cases there is no statistically significant difference between the property in the real network and in its 2.5k-random graphs. In that sense these graphs, that is, random graphs whose degree distribution, degree correlations, and degree-dependent clustering $\bar{c}(k)$ are as in the original network, capture many other important properties of the real network.

Some properties always converge. This is certainly true for the microscopic properties in Fig. 3, simply confirming that our dk -sampling algorithm operates correctly. But many properties that are not fixed by the dk -distributions converge as well. Neither the concentration of subgraphs of size 3 nor the distribution of

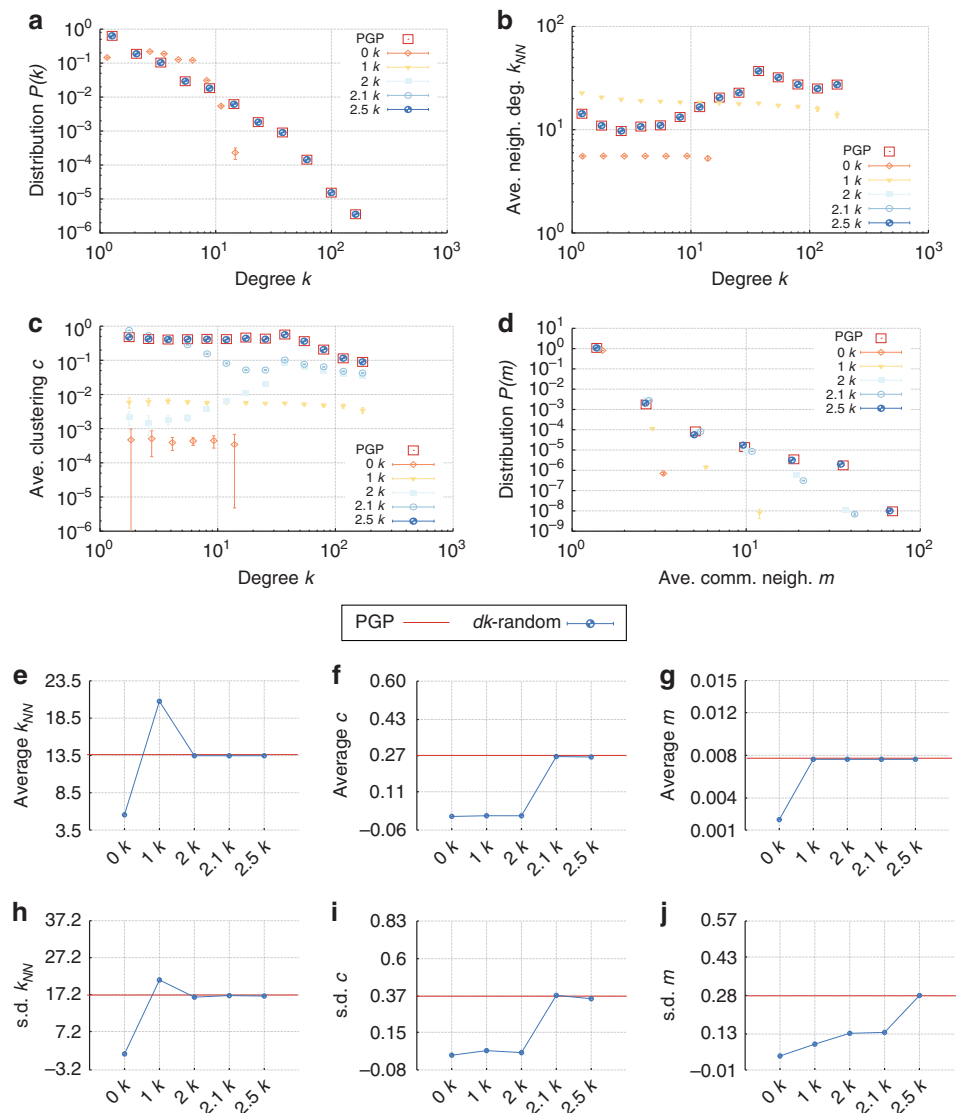


Figure 3 | Microscopic properties of the PGP network and its dk -random graphs. (a) The degree distribution $P(k)$, (b) average degree $\bar{k}_{nn}(k)$ of nearest neighbours of nodes of degree k , (c) average clustering $\bar{c}(k)$ of nodes of degree k , (d) the distribution $P(m)$ of the number m of common neighbours between all connected pairs of nodes, and (e–g) the means and (h–j) s.d. of the corresponding distributions. The error bars indicate the s.d. of the metrics across different graph realizations.

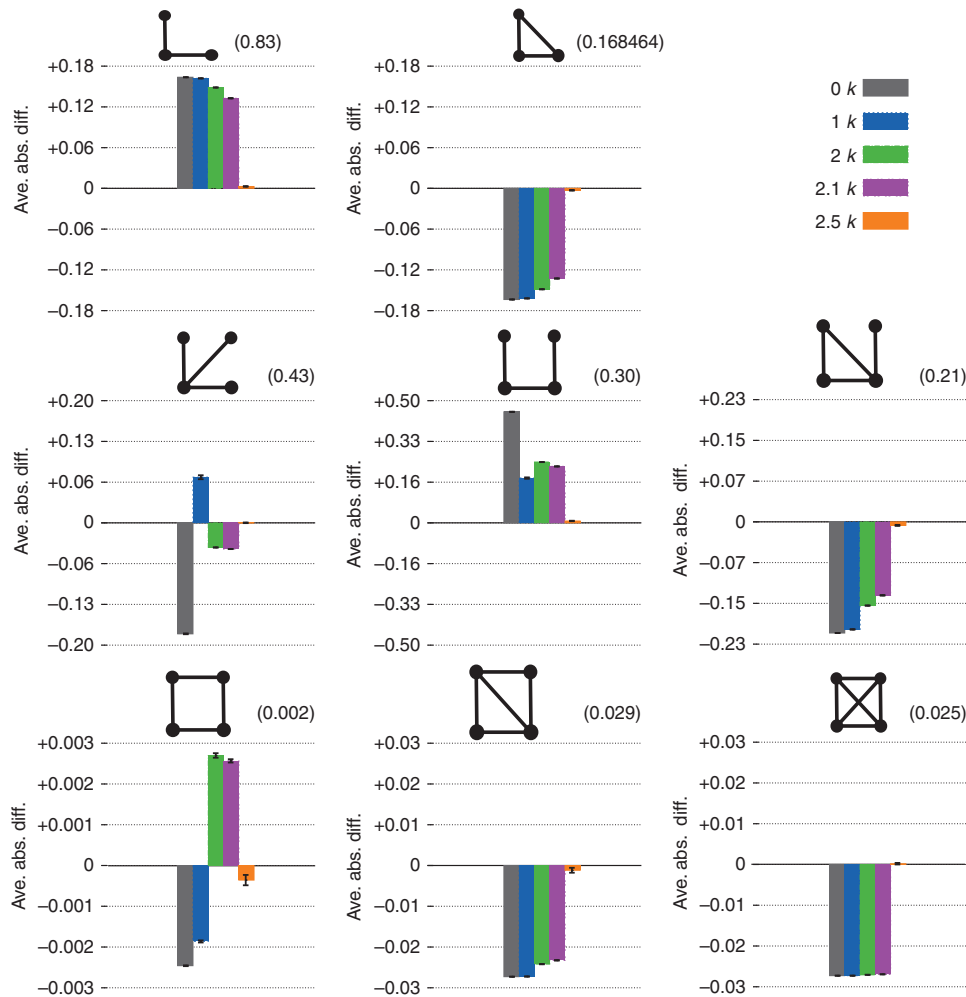


Figure 4 | The densities of subgraphs of size 3 and 4 in the PGP network and its dk -random graphs. The two different graphs of size 3 and six different graphs of size 4 are shown on each panel. The numbers on top of panels are the concentrations of the corresponding subgraph in the PGP network, while the histogram heights indicate the average absolute difference between the subgraph concentration in the dk -random graphs and its concentration in the PGP network. The subgraph concentration is the number of given subgraphs divided by the total number of subgraphs of the same size. The error bars are the s.d. across different graph realizations.

the number of neighbours common to a pair of nodes are fully fixed by dk -distributions with any $d < 3$ by definition, yet $2.5k$ -random graphs reproduce them well in all the considered networks. Most subgraphs of size 4 are also captured at $d = 2.5$ in most networks, even though $d = 3$ would not be enough to exactly reproduce the statistics of these subgraphs. We note that the improvement in subgraph concentrations at $d = 2.5$ compared with $d = 2.1$ is particularly striking, Fig. 4. The mesoscopic and especially macroscopic properties converge more slowly as expected. Nevertheless, quite surprisingly, both mesoscopic properties (k -coreness and k -density) and some macroscopic properties converge nicely in most cases. The k -coreness, k -density, and the spectral properties, for instance, converge at $d = 2.5$ in all the considered cases other than Internet’s Fiedler value. In some cases a property, even global one, converges for $d < 2.5$. Betweenness, for example, a global property, converges at $d = 1$ for the Internet and English word network.

Finally, there are ‘outlier’ networks and properties of poor or no dk -convergence. Many properties of the brain network, for example, exhibit slow or no convergence. We have also experimented with community structure inferred by different algorithms, and in most cases the convergence is either slow or non-existent as one could expect.

Discussion

In general, we should not expect non-local properties of networks to be exactly or even closely reproduced by random graphs with local constraints. The considered brain network is a good example of that this expectation is quite reasonable. The human brain consists of two relatively weakly connected parts, and no dk -randomization with low d is expected to reproduce this peculiar global feature, which likely has an impact on other global properties. And indeed we observe in Supplementary Note 3 that its two global properties, the shortest path distance and betweenness distributions, differ drastically between the brain and its dk -randomizations.

Another good example is community structure, which is not robust with respect to dk -randomizations in all the considered networks. In other words, dk -randomizations destroy the original peculiar cluster organization in real networks, which is not surprising, as clusters have too many complex non-local features such as variable densities of internal links, boundaries and so on, which dk -randomizations, even with high d , are expected to affect considerably.

Surprisingly, what happens for the brain and community structure does not appear representative for many other considered combinations of real networks and their properties.

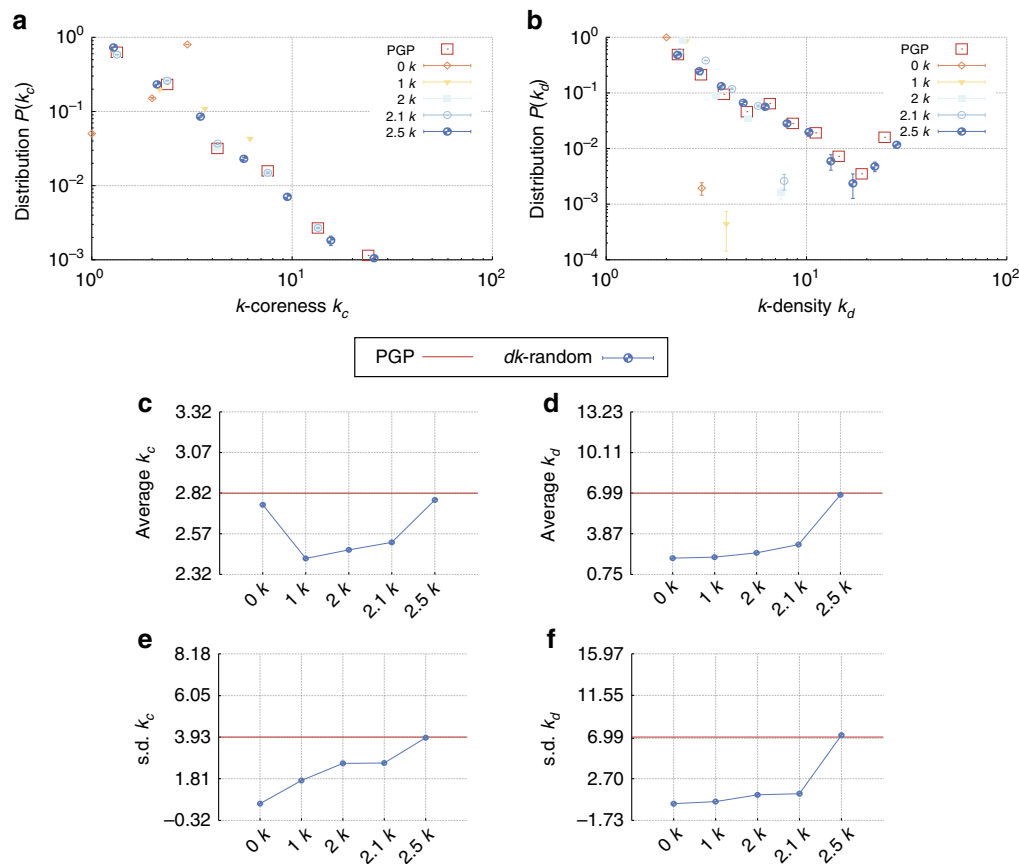


Figure 5 | Mesoscopic properties, the k -coreness and k -density distributions, in the PGP network and its dk -random graphs. The figure shows the distributions $P(k_{c,d})$ of (a) node k -coreness k_c and (b) edge k -density k_d , and their (c,d) means and (e,f) s.d. The k_c -core of a graph is its maximal subgraph in which all nodes have degree at least k_c . The k_d -core of a graph is its maximal subgraph in which all edges have multiplicity at least $k_d - 2$; the multiplicity of an edge is the number of common neighbours between the nodes that this edge connects, or equivalently the number of triangles that this edge belongs to. A node has k -coreness k_c if it belongs to the k_c -core but not to the $k_c + 1$ -core. An edge has k -density k_d if it belongs to the k_d -core but not to the $k_d + 1$ -core. The error bars indicate the s.d. of the metrics across different graph realizations.

As a possible explanation, one can think of constraint-based modelling as a satisfiability (SAT) problem: find the elements of the adjacency matrix (1/0, True/False) such that all the given constraints in terms of the functions of the marginals (degrees) of this matrix are obeyed. We then expect that the $3k$ -constraints already correspond to an NP-hard SAT problem, such as 3-SAT, with hardness coming from the global nature of the constraints in the problem. However, many real-world networks evolve based mostly on local dynamical rules and thus we would expect them to contain correlations with $d < 3$, that is, below the NP-hard barrier. The primate brain, however, has likely evolved through global constraints, as indicated by the dense connectivity across all functional areas and the existence of a strong core-periphery structure in which the core heavily concentrates on areas within the associative cortex, with connections to and from all the primary input and subcortical areas¹².

However, in most cases, the considered networks are dk -random with $d \leq 2.5$, that is, $d \leq 2.5$ is enough to reproduce not only basic microscopic (local) properties but also mesoscopic and even macroscopic (global) network properties^{6–10}. This finding means that these more sophisticated properties are effectively random in the considered networks, or more precisely, that the observed values of these properties are effective consequences of particular degree distributions and, optionally, degree correlations and clustering that the networks have. This further implies that attempts to find explanations for these

complex but effectively random properties should probably be abandoned, and redirected to explanations of why and how degree distributions, correlations and clustering emerge in real networks, for which there already exists a multitude of approaches^{52–57}. On the other hand, the features that we found non-random do require separate explanations, or perhaps a different system of null models.

We reiterate that the dk -randomization system makes it clear that there is no *a priori* preferred null model for network randomization. To tell how statistically significant a particular feature is, it is necessary to compare this feature in the real network against the same feature in an ensemble of random graphs, a null model. But one is free to choose any random graph model. In particular, any d defines a random graph ensemble, and we find that many properties, most notably the frequencies of small subgraphs that define motifs¹¹, strongly depend on d for many considered networks. Therefore, choosing any specific value of d , or more generally, any specific null model to study the statistical significance of a particular structural network feature, requires some non-trivial justification before this feature can be claimed important for any network function.

Yet another implication of our results is that if one looks for network topology generators that would veraciously reproduce certain properties of a given real network—a task that often comes up in as diverse disciplines as biology⁵⁸ and computer science⁵⁹—one should first check how dk -random these

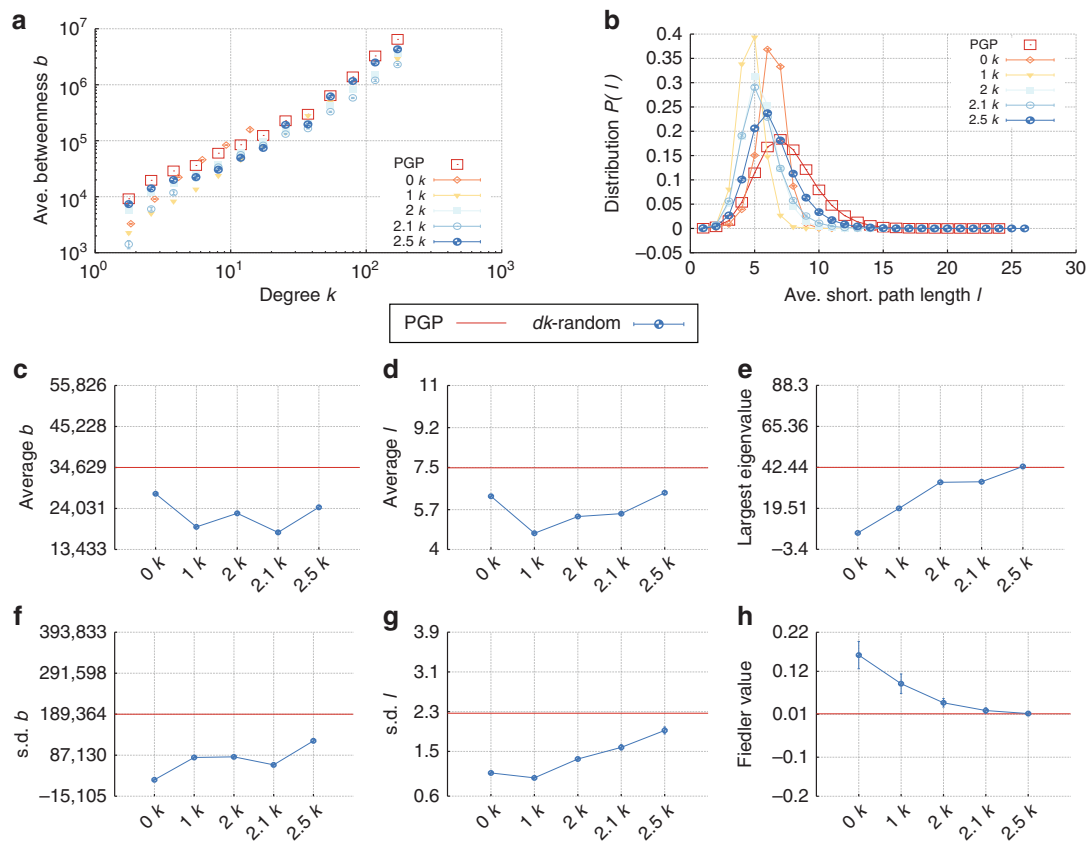


Figure 6 | Macroscopic properties of the PGP network and its dk -random graphs. (a) The average betweenness $\bar{b}(k)$ of nodes of degree k , (b) the distribution $P(l)$ of hop lengths l of the shortest paths between all pairs of nodes, the (c,d) means and (f,g) s.d. of the corresponding distributions, (e) the largest eigenvalues of the adjacency matrix A , and (h) the Fiedler value, which is the spectral gap (the second largest eigenvalue) of the graph's Laplacian matrix $L = D - A$, where D is the degree matrix, $D_{ij} = \delta_{ij}k_i$, δ_{ij} the Kronecker delta, and k_i the degree of node i . The error bars indicate the s.d. of the metrics across different graph realizations.

properties are. If they are dk -random with low d , then one may not need any sophisticated mission-specific topology generators. The dk -random graph-generation algorithms discussed here can be used for that purpose in this case. We note that there exists an extension of a subset of these algorithm for networks with arbitrary annotations of links and nodes⁶⁰—directed or coloured (multilayer) networks, for instance.

The main caveat of our approach is that we have no proof that our dk -random graph generation algorithms for $d=2.1$ and $d=2.5$ sample graphs uniformly at random from the ensemble. The random-graph ensembles and edge-rewiring processes employed here are known to suffer from problems such as degeneracy and hysteresis^{35,61,62}. Ideally, we would wish to calculate analytically the exact expected value of a given property in an ensemble. This is currently possible only for very simple properties in soft ensembles with $d=0, 1, 2$ (refs 36,37). Some mathematically rigorous results are available for $d=0, 1$ and for some exponential random-graph models^{28,34}. Many of these results rely on graphons²⁴ that are applicable to dense graphs only, while virtually all real networks are sparse⁴⁹. Some rigorous approaches to sparse networks are beginning to emerge^{63,64}, but the rigorous treatment of global properties, which tend to be highly non-trivial functions of adjacency matrices, in random graph ensembles with $d > 2$ constraints, appear to be well beyond the reach in the near future. Yet if we ever want to fully understand the relationship between the structure, function and dynamics of real networks, this future research direction appears to be of a paramount importance.

References

- Newman, M. E. J. *Networks: An Introduction* (Oxford Univ. Press, 2010).
- Barrat, A. *Dynamical Processes on Complex Networks* (Cambridge Univ. Press, 2008).
- Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**, 167–256 (2003).
- Adilson, E. Motter and Zoltán Toroczkai. Introduction: optimization in networks. *Chaos* **17**, 026101 (2007).
- Burda, Z., Krzywicki, A., Martin, O. C. & Zagorski, M. Motifs emerge from function in model gene regulatory networks. *Proc. Natl Acad. Sci. USA* **108**, 17263–17268 (2011).
- Vázquez, A. et al. The topological relationship between the large-scale attributes and local interaction patterns of complex networks. *Proc. Natl Acad. Sci. USA* **101**, 17940–17945 (2004).
- Guimerá, R., Sales-Pardo, M. & Luis, A. N. A. Classes of complex networks defined by role-to-role connectivity profiles. *Nat. Phys.* **3**, 63–69 (2007).
- Takemoto, K., Oosawa, C. & Akutsu, T. Structure of n-clique networks embedded in a complex network. *Phys. A* **380**, 665–672 (2007).
- Foster, D. V., Foster, J. G., Grassberger, P. & Paczuski, M. Clustering drives assortativity and community structure in ensembles of networks. *Phys. Rev. E* **84**, 066117 (2011).
- Colomer-de-Simón, P., Serrano, M. Á., Beiró, M. G., Alvarez-Hamelin, J. I. & Boguñá, M. Deciphering the global organization of clustering in real complex networks. *Sci. Rep.* **3**, 2517 (2013).
- Milo, R. et al. Network motifs: simple building blocks of complex networks. *Science* **298**, 824–827 (2002).
- Markov, N. T. et al. Cortical high-density counterstream architectures. *Science* **342**, 1238406 (2013).
- Luis, A. Amaral and Roger Guimera. Complex networks: Lies, damned lies and statistics. *Nat. Phys.* **2**, 75–76 (2006).
- Colizza, V., Flammini, A., Ángeles Serrano, M. & Vespignani, A. Detecting rich-club ordering in complex networks. *Nat. Phys.* **2**, 110–115 (2006).

15. Trevino, III J., Nyberg, A., Del Genio, C. I. & Bassler, K. E. Fast and accurate determination of modularity and its effect size. *J. Stat. Mech.* **15**, P02003 (2015).
16. Mahadevan, P., Krioukov, D., Fall, K. & Vahdat, A. Systematic topology analysis and generation using degree correlations. *Comput. Commun. Rev.* **36**, 135–146 (2006).
17. Yaveroglu, Ö. N. *et al.* Revealing the hidden language of complex networks. *Sci. Rep.* **4**, 4547 (2014).
18. Karrer, B. & Newman, M. E. J. Random graphs containing arbitrary distributions of subgraphs. *Phys. Rev. E* **82**, 066118 (2010).
19. Coolen, A. C. C., Fraternali, F., Annibale, A., Fernandes, L. & Kleijnung, J. *Modelling Biological Networks via Tailored Random Graphs* 309–329 (John Wiley & Sons, Ltd, 2011).
20. Coolen, A. C. C., Martino, A. & Annibale, A. Constrained Markovian Dynamics of Random Graphs. *J. Stat. Phys.* **136**, 1035–1067 (2009).
21. Annibale, A., Coolen, A. C. C., Fernandes, L., Fraternali, F. & Kleijnung, J. Tailored graph ensembles as proxies or null models for real networks I: tools for quantifying structure. *J. Phys. A Math. Gen.* **42**, 485001 (2009).
22. Roberts, E. S., Schlitt, T. & Coolen, A. C. C. Tailored graph ensembles as proxies or null models for real networks II: results on directed graphs. *J. Phys. A Math. Theor.* **44**, 275002 (2011).
23. Roberts, E. S. & Coolen, A. C. C. Unbiased degree-preserving randomization of directed binary networks. *Phys. Rev. E* **85**, 046103 (2012).
24. Lovász, L. *Large Networks and Graph Limits* (American Mathematical Society, 2012).
25. Erdős, P. & Rényi, A. On Random Graphs. *Publ. Math.* **6**, 290–297 (1959).
26. Bender, E. & Canfield, E. The asymptotic number of labelled graphs with given degree distribution. *J. Comb. Theor. A* **24**, 296–307 (1978).
27. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 26118 (2001).
28. Chatterjee, S., Diaconis, P. & Sly, A. Random graphs with a given degree sequence. *Ann. Appl. Probab.* **21**, 1400–1435 (2011).
29. Stanton, I. & Pinar, A. Constructing and sampling graphs with a prescribed joint degree distribution. *J. Exp. Algorithm.* **17**, 3.1 (2012).
30. Bianconi, G. The entropy of randomized network ensembles. *Eur. Lett.* **81**, 28005 (2008).
31. Barvinok, A. & Hartigan, J. A. The number of graphs and a random graph with a given degree sequence. *Random Struct. Algorithm.* **42**, 301–348 (2013).
32. Holland, P. W. & Leinhardt, S. An exponential family of probability distributions for directed graphs. *J. Am. Stat. Assoc.* **76**, 33–50 (1981).
33. Park, J. & Newman, M. E. J. Statistical mechanics of networks. *Phys. Rev. E* **70**, 66117 (2004).
34. Chatterjee, S. & Diaconis, P. Estimating and understanding exponential random graph models. *Ann. Stat.* **41**, 2428–2461 (2013).
35. Horvát, S., Czabarka, É. & Toroczkai, Z. Reducing degeneracy in maximum entropy models of networks. *Phys. Rev. Lett.* **114**, 158701 (2015).
36. Squartini, T. & Garlaschelli, D. Analytical maximum-likelihood method to detect patterns in real networks. *N. J. Phys.* **13**, 083001 (2011).
37. Squartini, T., Mastrandrea, R. & Garlaschelli, D. Unbiased sampling of network ensembles. *N. J. Phys.* **17**, 023052 (2015).
38. Maslov, S., Sneppen, K. & Alon, U. *Handbook of Graphs and Networks* chapter 8 (Wiley-VCH, 2003).
39. Maslov, S., Sneppen, K. & Zaliznyak, A. Detection of topological patterns in complex networks: Correlation profile of the Internet. *Phys. A* **333**, 529–540 (2004).
40. Gjoka, M., Kurant, M. & Markopoulou, A. In *2013 Proceedings of IEEE INFOCOM 1968–1976* (IEEE, 2013).
41. Kim, H., Toroczkai, Z., Erdős, P. L., Miklós, I. & Székely, L. A. Degree-based graph construction. *J. Phys. A Math. Theor.* **42**, 392001 (2009).
42. Del Genio, C. I., Kim, H., Toroczkai, Z. & Bassler, K. E. Efficient and exact sampling of simple graphs with given arbitrary degree sequence. *PLoS ONE* **5**, e10012 (2010).
43. Kim, H., Del Genio, C. I., Bassler, K. E. & Toroczkai, Z. Constructing and sampling directed graphs with given degree sequences. *N. J. Phys.* **14**, 023012 (2012).
44. Bassler, K. E., Del Genio, C. I., Erdős, P. L., Miklós, I. & Toroczkai, Z. Exact sampling of graphs with prescribed degree correlations. *N. J. Phys.* **17**, 083052 (2015).
45. Zlatic, V. *et al.* On the rich-club effect in dense and weighted networks. *Eur. Phys. J. B* **67**, 271–275 (2009).
46. Czabarka, É., Dutle, A., Erdős, P. L. & Miklós, I. On realizations of a joint degree matrix. *Discret. Appl. Math.* **181**, 283–288 (2015).
47. Jamakovic, A., Mahadevan, P., Vahdat, A., Boguñá, M. & Krioukov, D. How small are building blocks of complex networks. Preprint at <http://arxiv.org/abs/0908.1143> (2009).
48. Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J. & Alon, U. On the uniform generation of random graphs with prescribed degree sequences. Preprint at <http://arxiv.org/abs/cond-mat/0312028> (2003).
49. Del Genio, C., Gross, T. & Bassler, K. E. All scale-free networks are sparse. *Phys. Rev. Lett.* **107**, 1–4 (2011).
50. Gjoka, M., Tillman, B. & Markopoulou, A. In *2015 Proceedings of IEEE INFOCOM*. pages 1553–1561 (IEEE, 2015).
51. de Simon, P. C. RandNetGen: a Random Network Generator. URL: <http://polcolomer.github.io/RandNetGen/>.
52. Dorogovtsev, S. N., Mendes, J. & Samukhin, A. Size-dependent degree distribution of a scale-free growing network. *Phys. Rev. E* **63**, 062101 (2001).
53. Klemm, K. & Eguluz, V. Highly clustered scale-free networks. *Phys. Rev. E* **65**, 036123 (2002).
54. Vázquez, A. Growing network with local rules: Preferential attachment, clustering hierarchy, and degree correlations. *Phys. Rev. E* **67**, 056104 (2003).
55. Serrano, M. Á. & Baguna, M. Tuning clustering in random networks with arbitrary degree distributions. *Phys. Rev. E* **72**, 036133 (2005).
56. Papadopoulos, F., Kitsak, M., Serrano, M. Á., Boguñá, M. & Krioukov, D. Popularity versus similarity in growing networks. *Nature* **489**, 537–540 (2012).
57. Bianconi, G., Darst, R. K., Iacovacci, J. & Fortunato, S. Triadic closure as a basic generating mechanism of communities in complex networks. *Phys. Rev. E* **90**, 042806 (2014).
58. Kuo, P., Banzhaf, W. & Leier, A. Network topology and the evolution of dynamics in an artificial genetic regulatory network model created by whole genome duplication and divergence. *Biosystems* **85**, 177–200 (2006).
59. Medina, A., Lakhina, A., Matta, I. & Byers, J. In *MASCOTS 2001, Proceedings of the Ninth International Symposium in Modeling, Analysis and Simulation of Computer and Telecommunication Systems* 346–353 (Washington, DC, USA, 2001).
60. Dimitropoulos, X., Krioukov, D., Riley, G. & Vahdat, A. Graph annotations in modeling complex network topologies. *ACM T Model. Comput. S* **19**, 17 (2009).
61. Foster, D., Foster, J., Paczowski, M. & Grassberger, P. Communities, clustering phase transitions, and hysteresis: Pitfalls in constructing network ensembles. *Phys. Rev. E* **81**, 046115 (2010).
62. Roberts, E. S. & Coolen, A. C. C. Random graph ensembles with many short loops. In *ESAIM Proc. Surv.* **47**, 97–115 (2014).
63. Bollobás, B. & Riordan, O. Sparse graphs: metrics and random models. *Random Struct. Algorithm.* **39**, 1–38 (2011).
64. Borgs, C., Chayes, J. T., Cohn, H. & Zhao, Y. An L^p theory of sparse graph convergence I: Limits, sparse random graph models, and power law distributions. Preprint at <http://arxiv.org/abs/1401.2906> (2014).
65. Beiró, M. G., Alvarez-Hamelin, J. I. & Busch, J. R. A low complexity visualization tool that helps to perform complex systems analysis. *N. J. Phys.* **10**, 125003 (2008).

Acknowledgements

We acknowledge financial support by NSF Grants No. CNS-1039646, CNS-1345286, CNS-0722070, CNS-0964236, CNS-1441828, CNS-1344289, CNS-1442999, CCF-1212778, and DMR-1206839; by AFOSR and DARPA Grants No. HR0011-12-1-0012 and FA9550-12-1-0405; by DTRA Grant No. HDTRA-1-09-1-0039; by Cisco Systems; by the Ministry of Education, Science, and Technological Development of the Republic of Serbia under Project No. ON171017; by the ICREA Academia Prize, funded by the *Generalitat de Catalunya*; by the Spanish MINECO Project No. FIS2013-47282-C2-1-P; by the *Generalitat de Catalunya* Grant No. 2014SGR608; and by European Commission Multiplex FP7 Project No. 317532.

Author contributions

All authors contributed to the development and/or implementation of the concept, discussed and analysed the results. C.O., M.M.D., and P.C.S. implemented the software for generating dk -graphs and analysed their properties. D.K. wrote the manuscript, incorporating comments and contributions from all authors.

Additional information

Supplementary Information accompanies this paper at <http://www.nature.com/naturecommunications>

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at <http://npg.nature.com/reprintsandpermissions>.

How to cite this article: Orsini, C. *et al.* Quantifying randomness in real networks. *Nat. Commun.* 6:8627 doi: 10.1038/ncomms9627 (2015).



This work is licensed under a Creative Commons Attribution 4.0 International License. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in the credit line; if the material is not included under the Creative Commons license, users will need to obtain permission from the license holder to reproduce the material. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>

RESEARCH ARTICLE

A Theoretical Model for the Associative Nature of Conference Participation

Jelena Smiljanić^{1,2}, Arnab Chatterjee³, Tomi Kauppinen⁴, Marija Mitrović Dankulov^{1*}

1 Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia, **2** School of Electrical Engineering, University of Belgrade, P.O. Box 35-54, 11120 Belgrade, Serbia, **3** Condensed Matter Physics Division, Saha Institute of Nuclear Physics, 1/AF Bidhannagar, Kolkata 700064, India, **4** Aalto University School of Science, P.O. Box 11000, FI-00076 AALTO, Finland

* mitrovic@ipb.ac.rs



OPEN ACCESS

Citation: Smiljanić J, Chatterjee A, Kauppinen T, Mitrović Dankulov M (2016) A Theoretical Model for the Associative Nature of Conference Participation. PLoS ONE 11(2): e0148528. doi:10.1371/journal.pone.0148528

Editor: Matjaz Perc, University of Maribor, SLOVENIA

Received: November 26, 2015

Accepted: December 8, 2015

Published: February 9, 2016

Copyright: © 2016 Smiljanić et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data files are available from https://figshare.com/articles/Conference_Datasets/2066907.

Funding: This work was supported by ON171017, The Ministry of Education, Science, and Technological Development of the Republic of Serbia (<http://www.mpn.gov.rs/>: J.S. and M.M.D.); and 675121, The European Commission (http://ec.europa.eu/index_en.htm: M.M.D.).

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Participation in conferences is an important part of every scientific career. Conferences provide an opportunity for a fast dissemination of latest results, discussion and exchange of ideas, and broadening of scientists' collaboration network. The decision to participate in a conference depends on several factors like the location, cost, popularity of keynote speakers, and the scientist's association with the community. Here we discuss and formulate the problem of discovering how a scientist's previous participation affects her/his future participations in the same conference series. We develop a stochastic model to examine scientists' participation patterns in conferences and compare our model with data from six conferences across various scientific fields and communities. Our model shows that the probability for a scientist to participate in a given conference series strongly depends on the balance between the number of participations and non-participations during his/her early connections with the community. An active participation in a conference series strengthens the scientist's association with that particular conference community and thus increases the probability of future participations.

Introduction

Social data at a large scale is nowadays available over the internet. Researchers are making the best use of these data to find trends, statistics and patterns, which sometime reveal as robust features, similar to 'laws' in natural science. In recent years, a huge community of researchers [1] including mathematicians, statisticians, computer scientists, theoretical physicists, sociologists, economists, financial analysts, geographers, anthropologists, and biologists of various sub-disciplines have contributed to a larger, developing field, commonly known as 'computational social science' [2]. Empirical data, after a rigorous analysis produces information that is of immense interest for theoreticians. Statistical mechanics, which has been proved to be versatile in modeling phenomena across different areas of physics, and beyond, seems to be the most desired tool even for the above emerging discipline [3, 4].

The abundance of a new data about scientific activities such as publications, collaborations, and citations led to the emergence of a new interdisciplinary field of research about science and how science works [5]. These studies provide insights about the impact of scientists and their publications [6–8], authors' reputation and scientific success [9], patterns of collaboration and their impact on authors' reputation [10, 11], the role of cumulative advantage in career longevity [12, 13] and scientific mobility [14] among many other things. Despite the attention given to publication records and citation patterns, another integral part of modern science, scientific meetings, have so far been largely overlooked. This negligence is particularly interesting, given the pervasive role of the meetings in scientific disciplines. Scientific meetings provide arenas for a fast dissemination of the latest results, exchange and evaluation of ideas as well as a knowledge extension. However, the most important function of scientific meetings is to facilitate social contacts. They provide an opportunity and platform to extend the network of collaborators through the creation of new contacts, and to strengthen existing links by getting reacquainted with old friends.

Undoubtedly, conference participation has a very positive impact on scientific career. In addition to the opportunities they provide, attending a scientific meeting can be very costly, both in terms of time and money. Bearing in mind that the number of national and international meetings have drastically increased in the last few decades, it is clear that scientists are now pressed to make a careful selection of the meetings they will attend. Extensive studies [15–17] have shown that conference characteristics, such as the attractiveness and the reachability of the location or the choice of keynote speakers affect the decision of scientists to attend a meeting. The role of the social component in conference choice is so far unexplored, mainly due to lack of quality data. The social component, such as the association with a conference community or conference inclusiveness, are of crucial importance when it comes to whether a conference participation was beneficial or not. This is particularly evident in the case of young scientists, who are new to a community and struggle to overcome the social obstacle of an initial contact [18, 19]. One of the rare studies on conference participation [20] has shown that conferences have a stable core of regularly attending participants, regardless of the conference location and distance. Having in mind that characteristics like the attractiveness of a location and the quality of keynote speakers are fluctuating from one year to another, it is clear that social component of a conference strongly influence the scientists decision to attend the conference and their long-term participation patterns, accordingly.

The association with a conference community and conference inclusiveness, can have a strong influence on scientists persistence in participating at the specific conference. The problem of the order-parameter persistence (first-passage time), is a well studied phenomenon in non-equilibrium statistical dynamics in condensed matter systems [21]. Persistence is defined as the probability that fluctuating variable does not change the sign until time t , and for many non-equilibrium systems this probability decays with time as a power-law [21]. Here we carry out the analysis of persistence of participation patterns of more than 100000 scientists at six national and international conferences of different sizes and from different fields of science. We study the probability of total and successive number of participations, as well as the distribution of time lags between two successive participations. We find that all three measured probabilities have a shape of a truncated power law, regardless of the conference size and degree of specialization. This indicates that the probability for a participant to attend the next meeting is not constant, but rather it grows/decays with a number of participations/non-participations. This observation is directly related to the strength of the association with the conference community. We propose a microscopic stochastic model which includes this influence of balance between the number of participations and non-participations, as well as the role of conference inclusiveness, on the probability to attend the conference next year. Results of our

model show that the studied conferences have a relatively low inclusiveness, i.e. the probability for a scientist to participate in the next meeting after the first attendance. We also show that conference attendance is characterized by *positive feedback*. The growth in the total number of participations results in a stronger attractiveness of the conference community to participants, and vice versa. Longevity of scientific career of publishing in scientific journals is also characterized by a power-law distribution with an exponential cut-off [12]. Using the empirical analysis and stochastic model Petersen et al. [12] have shown that longevity and past success of scientists lead to cumulative advantage in further development of their career. Although the distribution of career longevity and conference persistence have a similar behaviour, there is a significant difference of characteristic exponents, which indicates that a different mechanism underlie these two phenomena.

This paper is structured as follows: first, we perform empirical analysis of participation patterns for six conferences. We then propose and describe the model of conference participation dynamics. Finally, we perform numerical simulations and discuss some properties of the model, and estimate the values of parameters that correspond to empirical data.

Results

Data set

For our empirical analysis we use data for six conference series in different fields of science. We collected and filtered information about abstracts presented at the American Physical Society March Meeting (APSMM), American Physical Society April Meeting (APSAM), Society for Industrial and Applied Mathematics Annual Meetings (SIAM), Neural Information Processing Systems Conference (NIPS), International Conference on Supercomputing (ICS) and Annual International Conference on Research in Computational Molecular Biology (RECOMB). All these scientific meetings are held annually, but they differ in the topic, sizes, degree of specialisation, longevity and degree of localisation (national versus international). When it comes to the meeting size it can vary from a few dozens, like ICS and RECOMB, to several thousands of participants at APSMM. Some of these meetings are on highly focused topic, NIPS, while others are designed to cover the entire scientific fields, like APSMM, APSAM and SIAM. Four of these conferences (SIAM, NIPS, ICS and RECOMB) have an international character with venues all over the world, while APSMM and APSAM are annual conferences of American Physical Society which are always held in North American cities. APSMM, SIAM and APSAM are conferences with a long tradition, while first meetings of NIPS, ICS and RECOMB have been organized during late 80s and early 90s. Detailed information about conferences and data is given in [S1 File](#).

To be able to track participants at the conference over the years, we have labeled them based on name, affiliation and co-authors and performed author name disambiguation (see [Methods](#) for details). We are interested in studying the participation patterns of scientists starting from their first attendance at the conference series. Thus, for conferences for which the data are not available from their beginning (APSMM, APSAM and SIAM), we have filtered out the authors that may have attended the conference before the starting year in our dataset (see [Methods](#) for the details of our filtering procedure).

Empirical results

For all scientists we have the information about the years of their appearance as authors in the book of abstracts of particular a conference series. The information about the list of authors who actually attended the conference is not available for the conferences considered in this paper. Hence, as a proxy for a conference participation in a given year, we use the appearance

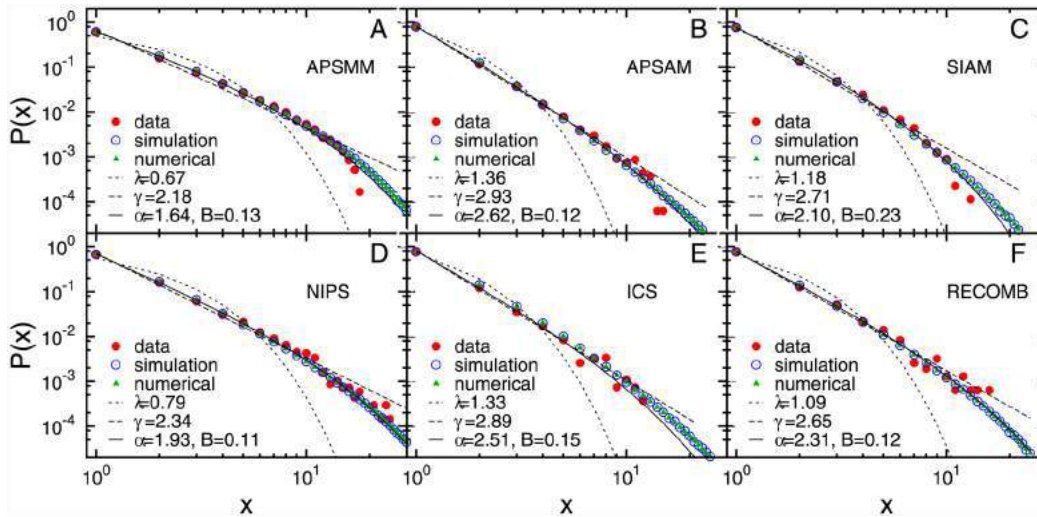


Fig 1. The total number of participations. The probability distribution of the total number of participations obtained from the empirical data (red circles), simulations (blue circles) and numerical iterative algorithm (green triangles). The full line is the best fit to truncated power law, $x^{-\alpha} e^{-Bx}$, while the dashed and dash-dot line denote the best fit to power-law distribution, $x^{-\gamma}$ and exponential distribution, $e^{-\lambda x}$, respectively.

doi:10.1371/journal.pone.0148528.g001

of a scientist as a co-author of at least one abstract in conference proceeding for that year. Not all authors that are mentioned in the book of abstracts have actually attended the conference, but one can argue that as co-authors they have actively contributed to the material presented and thus participate as a contributors in the conference [15].

First we analyse *the total number of author’s participations* (the number of times an author has participated), x , at the given conference series. Fig 1, shows the probability distribution of the total number of participations, $P(x)$, averaged over all participants, for each of the six analysed conferences. The comparison of the quality of fits between exponential, power-law and truncated power-law, Fig 1, shows that all curves are very well represented by power law with exponential cut-off (see Methods), with the value of exponent $\alpha \in (1.6, 2.7)$. The disparity in the total number of participations indicates that most scientists belong to the group of occasional participants, with more than half of all participants attending a particular conference only once. For instance, the percentage of all participants that attend the conference only once is the highest for APSAM and ICS, around 81%, and the lowest for APSMM and NIPS, 63% and 68% respectively. This observation indicates that communities of all these conferences have a relatively low inclusiveness. On the other hand, it is also clear that some of the participants are very regular, attending the conference (almost) every year. These participants form the group of regular attendees whose conference participation is mainly driven by social factors, i.e. their sense of *association with the community*.

In the case of when the probability to attend a conference is constant or random, the expected distribution of total number of attendances is of exponential type. Thus, the power-law nature of the distribution of total participations strongly suggests that the probability of participation at some future conference increases with the number of previous participations. By participating frequently at a particular conference scientists not only expand, but also strengthen, their collaboration network which leads to their further engagement with the community.

We further explore the participation patterns by analysing the number of successive participations (Fig 2) and the time lag between two successive participations (Fig 3). The distributions of these quantities also exhibit the truncated power-law behaviour (see Methods). The observed

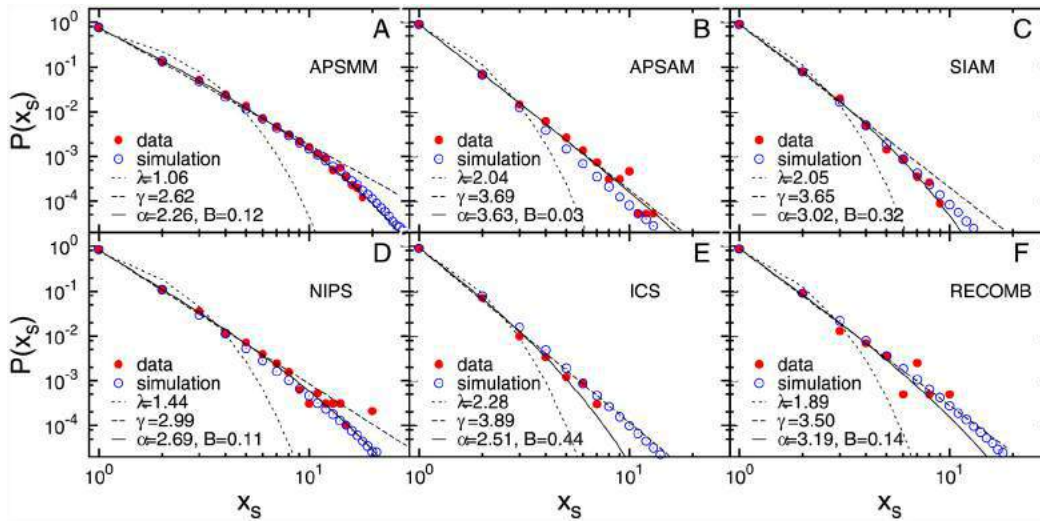


Fig 2. The number of successive participations. The probability distribution of the number of successive participations, x_s , obtained from empirical data (red circles) and numerical simulations of the model (blue circles). The full, dashed and dash-dot line are the best fit to truncated power law, power-law and exponential function respectively.

doi:10.1371/journal.pone.0148528.g002

distributions of the number of successive participations, with exponent $2 \leq \alpha \leq 4$, suggests that even frequent attendees make a pause in their participation, although these breaks are usually short, i.e. long breaks of five and more years occur with a low probability, Fig 3. A long-period of non-participation results in fading of existing collaboration ties with the community while new ones are never formed. Due to this fading, the probability to attend the meeting decreases with total number of non-participations. This indicates that conference participation of most scientists takes place in a limited period of time with a relatively short and small number of breaks.

As it was shown in Ref [12] the distribution of the journal career longevity exhibits a truncated power-law behaviour with cut-off around 10 years. The exponential cut-off in the distribution of all three measures is a consequence of the two combined finite-size effects that influence the asymptotic behaviour, the finite life time of scientist’s association with one community or her/his career in one field of research or in science in general [12], and limitations of used datasets. This effect will be also observed in the distribution of conference participations. The end of a career inevitably results in a termination of participation in conferences and thus also the conference community membership. Also, used datasets have a relatively short time span (less than three decades), due to which they do not include scientists with long careers [12]. Both of these effects affect the value of the exponential cut-off, which is lower in the case of conference participation, between 4 and 9 years, compared to the one observed for the career longevity.

Model

The empirical results from six different series shown in the previous section indicate that the probability for a scientist to attend the next meeting of a conference series depends on the balance of previous participations and non-participations. Petersen et al. [12] show that Matthew (*rich get richer*) effect is responsible for the career longevity in several competitive professions, including science. They argue that it becomes easier to move forward in the career with an increasing past success of an individual, and show, using their stochastic career progressive model, that this mechanism leads to a truncated power-law distribution of the career longevity.

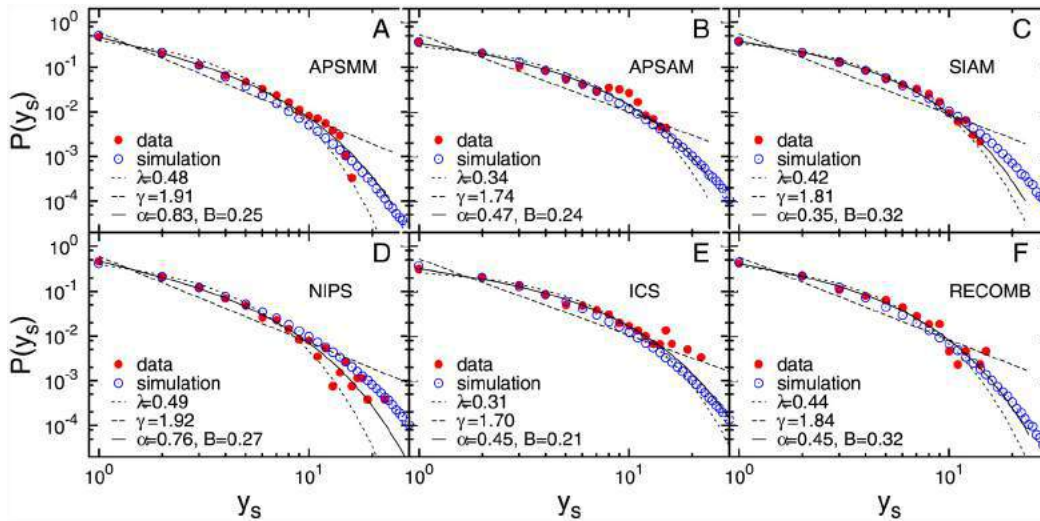


Fig 3. The time lag between the two successive participations. The probability distribution of the time lags between two consecutive conference participations y_s : empirical data (red circles) and numerical simulations data (blue circles). The lines correspond to respective fits as in Figs 1 and 2.

doi:10.1371/journal.pone.0148528.g003

In their model, they assume that the stochastic process governing career progress is similar to Poisson process, where progress is made at any given step with the rate $g(x) \equiv 1 - \exp[-(x/x_c)^\alpha]$, where $1/x_c$ is a hazard rate corresponding to random career ending while the parameter α is the same as power-law exponent in the pdf of career longevity. Using this model for $\alpha < 1$ they were able to obtain truncated power-law distributions for career duration in several professions.

The empirical results of conference participation patterns suggest that the probability for a scientist to participate in a conference is not constant or random, but that it rather grows with the number of participations. This is reflected in the increase of proportion of authors who are going to attend the conference next year with total number of previous conference attendance (see Figure A in S1 File). Higher number of participations of a scientist at the conference results in better connections with the community and thus higher probability that the author will participate in the following conference. But unlike career longevity, where the length of the waiting times between two successive steps in the career does not influence the progress rate, the probability for conference participation is strongly influenced by the number and length of pauses (Figure B in S1 File). The longer the scientists are absent from the community the weaker are their connections and lower are the probabilities to participate in the following events. For this reason and the fact that the pdf obtained from the model proposed in Ref [12] exhibits a truncated power-law only for the exponents $\alpha < 1$ Petersen et al. model [12] cannot be applied for modelling conference participation dynamics.

We propose a new stochastic model for conference attendance dynamics which can explain our empirical findings. Our model is based on a 2-bin generalized Pólya process [22–24] and random termination time of a career. As opposed to the Petersen model where the progress rate depends only on the current position of scientist in his/her career, the 2-bin generalized Pólya incorporates dependence on the balance between participations and non-participations. Let x stands for the total number of participations at the conference, y stands for the number of conferences an author has not participated since she/he appeared at the conference for the first time and t is the number of events held, $t = x + y$. All authors start with $x = 1$ and $y = 0$. According to our model, the probability that a scientist with x total number of participations and y

number of non-participations will appear at the next conference is given by

$$g(x, y) = \frac{x^p}{x^p + (y + y_0)^p} = \frac{z^p}{1 + z^p}, \tag{1}$$

where $z = \frac{x}{y+y_0}$ measures the balance between participations and non-participations, parameter p is the *exponent* of the model, and y_0 determines the initial balance value. The probability that a scientist will not attend the next conference is equal to $1 - g(x, y)$. Depending on the exponent p , the function g can correspond to positive ($p > 1$) or negative feedback ($p < 1$) [22]. When $p = 1$ and $y_0 = 0$, the Eq 1 is equivalent to the equation for a Pólya-Eggenberg problem [25]. As we shall see in the following section, the value of the parameter p for all conferences is larger than one, suggesting that the conference participation dynamics is characterized by the *positive feedback*: scientists who participate in the conference frequently and make less and shorter pauses have a stronger association with the conference community and thus have a higher probability to participate in the following events. The value of the parameter y_0 determines the probability of a scientist to attend the next event after her/his first occurrence at the conference. According to our model this parameter is the same for all scientists attending one conference series, thus it can be interpreted as a measure of the conference community inclusiveness.

Evolution equation. The probability $P(x, t)$ for the author to have x conference participations after t conferences since his/her first participation is equal to the probability to attend the next conference $g(x - 1, t - x)$ times the probability of already attending $x - 1$ conferences at time $t - 1$ plus the probability of skipping the next conference $1 - g(x, t - 1 - x)$ times the probability of already attending x conferences at time $t - 1$:

$$P(x, t) = \frac{(x - 1)^p}{(x - 1)^p + (t - x + y_0)^p} P(x - 1, t - 1) + \frac{(t - 1 - x + y_0)^p}{x^p + (t - 1 - x + y_0)^p} P(x, t - 1). \tag{2}$$

The probability distribution $P(x)$ of the number of total conference attendances for a particular conference series is obtained by summing $P(x, t = T)$ over all possible T :

$$P(x) = \sum_{T=1}^{\infty} P(x, t = T) P(T), \tag{3}$$

where T denotes the duration of a scientist’s membership in the community. In our case, we assume that the duration of a scientist’s membership in a conference community can be terminated at any year after his/her first appearance with probability H , which gives the distribution of time intervals

$$P(T) = H(1 - H)^{T-1}. \tag{4}$$

Numerical simulation results

Since the analytical solution of Eq 3 cannot be obtained, we estimate the model parameters y_0 , H and p using numerical simulations (see [Methods](#)). The best estimates of the model parameters for each of the six conferences are given in Table G in [S1 File](#). As shown in Figs 1, 2 and 3, the model with the properly chosen parameters nicely reproduces the behaviour of participants at six conferences, for all three measured quantities.

For all six conferences the estimated value of parameter p is greater than 1, which suggests that the positive feedback mechanism underlies the conference participation dynamics. This means that the probability for a scientist to attend the next year event grows superlinearly with the balance between the number of participations and pauses (z). The value of the parameter y_0

together with the value of p determines the probability for a scientist to participate in the conference next year after his/her first participation, i.e. the initial inclusiveness of the conference community. Table H in [S1 File](#) shows the estimated value of the initial inclusiveness for all six conferences. The APSMM has the highest probability, around 25%, for newcomers to attend the conference next year, while APSAM has the lowest, 9%. One could assume that the size and diversity of topics of a conference have an essential influence on conference inclusiveness, but according to our results this is not the case. The ordering of the conferences according to size, Table H in [S1 File](#), and their initial inclusiveness do not correlate. APSAM is the second largest conference but has the lowest inclusiveness, while the RECOMB as the smallest conference is ranked as third and has the inclusiveness of 15%. Further, it follows from our results that the diversity of topics covered by the conference does not have a significant effect on the return probability of newcomers. Although the first ranked conference according to inclusiveness, APSMM, covers the widest range of topics among considered conferences, the APSAM and SIAM, which are also considered general conferences, have a lower inclusiveness than NIPS and RECOMB. This suggests that the conference inclusiveness is influenced by some other factors, which are not related to the size, degree of specialisation or localisation (national and international), but rather to social structure and openness of the conference community toward newcomers.

We solve [Eq 3](#) numerically using an iterative method (see SI for more details) and compare it with simulation results. [Fig 1](#) shows an excellent matching between results obtained using the iterative algorithm and numerical simulations for the estimated values of parameters.

Discussion and conclusion

The goal of this paper has been to investigate the conference participation patterns and propose a simple stochastic model of conference participation dynamics. The motivation behind this is to better understand the mechanisms that underlie the repeated participation in the same conference series and explore whether the conference series topic, size, degree of specialisation, longevity and degree of localisation (national and international) influence the participation probability and inclusiveness of the specific community. Our study is based on empirical analysis and modelling of authors participation at six different conference series in the last three decades: APSMM, APSAM, SIAM, NISP, ICS and RECOMB. We note here that it would be important to verify our findings with the data from other conferences.

The set of considered conferences is very heterogeneous. Although they differ in size, topic and topic diversity, national structure of participants and conference longevity, they are characterized with similar participation patterns. The distributions of the total number of participations for all six conferences exhibit the same, truncated power-law, behaviour with values of exponent α between 1.6 and 2.7. A similar behaviour is also observed for the distributions of the number of successive participations and the duration of pauses between them. The observed statistical evidence strongly imply that the dynamics of conference participation is governed by universal forces which are independent of the specific conference features or the scientific field. This and the fact that conferences often have a stable core of attending participants [20] suggests that these have social origins and that social factors, such as the association with a conference community and its inclusiveness, strongly influence the probability for a scientist to attend the future meetings and their participation patterns at the specific conference series, accordingly.

The observed truncated power-law behaviour of the distributions of participations indicates that the probability for a scientist to participate in the next year conference is growing (decreasing) with the balance between the number of participations and pauses. To further explore this

we proposed a stochastic model based on 2-bin generalized Pólya process which incorporates the dependence on the ratio between number of participations and pauses. Our model shows that the positive feedback mechanism underlies the conference participation dynamics. The probability for a scientist to attend a conference grows superlineary with the number of participations, while the frequent pauses have the opposite effect. The scientists who are able to overcome the initial obstacles and create social ties with the conference community by frequent participation at the beginning have a higher probability to attend the conference in the following years. A frequent participation strengthens the scientist's association with a conference community which further increases the probability for future participations. On the other hand, scientists with a small number of initial participations have a low probability to participate in the following conference, thus small number of participations, and eventually stop attending the conference. The initial inclusiveness of the specific conference community has the main influence on early participation patterns. As we showed, this inclusiveness does not depend on the size, degree of specialisation or topic of the conference, but rather on the openness of the community toward newcomers.

Our analysis indicates that social factors, such as the association with the community and the community inclusiveness are the main driving forces of conference participation dynamics. In general the community/group cohesion and the ability to attract and retain newcomers and other members influence the dynamics of their participation in group activities [26]. On the other hand, a member's engagement in group activities strengthens ties to other group/community members, and contributes to the creation of the bonding capital, while the ties of non-attendees dissolve and weaken with time [27]. Conference communities are just one example of these systems, thus we expect to observe the similar group participation patterns in other types of social communities, both online and offline. Further investigations and studies of other social systems will reveal and characterize the connection between a social network structure and group inclusiveness, and participation dynamics in group activities.

Methods

Data filtering Identification of the different authors may involve a few issues. On one hand, an author may use different spelling variants to sign his first and middle name. On the other hand, the author's name may be related to several different authors, thus using only the initials of the last name and first name increases additionally error rates in disambiguating the author names. In our data sets, data from NIPS and RECOMB conferences did not require additional cleaning, while for the SIAM and ICS data, we have used python fuzzy partial string matching of author's first and middle names, which gave a high accuracy. For APSMM and APSAM conferences, where data are highly heterogeneous, we have used a method described in [28] to disambiguate the author names. This method considers pairs of names that match on last name and first name initials. Then it groups the authors based on their affiliation and co-authors. Because the same affiliation could be formatted differently, the two affiliations were considered the same if their fuzzy token set ratio was higher than 50%.

The sources and detailed description of the data are given in Tables A, B and C in [S1 File](#). For NIPS, ICS and RECOMB, we have complete data from their very beginning. Remaining data sets required filtering out the authors with a high probability of attending conference before the starting year in our dataset, Y_0 . Therefore, for APSMM, APSAM and SIAM we have isolated authors with the first recorded year of conference attendance, smaller than $Y_0 + \langle \tau \rangle$, where $\langle \tau \rangle$ is the average waiting time between a consecutive conference attendance for all the authors who took part at the conference during the $[Y_0, Y_f]$ period. This way we excluded between 10% (APSMM and SIAM) and 25% (APSAM) authors from our analysis.

Functional fits We have used the maximum-likelihood fitting method [29] to fit three different functions to the probability distributions of the total number of participations, the number of and the time lags between two successive participations: exponential function $e^{-\lambda x}$, power-law function $x^{-\gamma}$ and truncated power-law $x^{-\alpha} e^{-Bx}$. It follows from the comparison of fits of these three functions to empirical data that the truncated power-law is the best fit for the probability distribution of all three measured quantities, see Figs 1, 2 and 3. In order to compare these three fits we calculate the log likelihood ratio, \mathcal{R} , and π -value (see Ref [29]) which compares the fits to the power-law with exponential cut-off with the pure power-law for the distribution of total number of participations (Table D in S1 File) and the number of successive participations (Table E in S1 File). In the case of nested distributions, the negative value of \mathcal{R} indicates that the larger family of distributions, in this case the truncated power-law, is a superior model. When the value of \mathcal{R} tends to 0, one can use π -value. The small π -value suggests that the smaller family of distributions, in this case power-law, can be ruled-out. Both the log likelihood ratio and the π -value indicate that the truncated power-law is a superior model compared to pure power-law for both distributions. A similar procedure can be applied for the comparison between truncated power-law and exponential fits, but since from the visual inspection it is clear that the distributions do not follow the exponential fits, we have omitted these results. The comparison between exponential and the power-law with exponential cut-off fit, given in Table F in S1 File, indicates that the power-law distribution with exponential cut-off fit is better than exponential fit for the distribution of the time lags. For all six conferences, the power-law with exponential cut-off distribution gives the best fit for all three empirical distributions.

Parameter estimation We simulate the model for $N = 100000$ different authors. Starting from $x = 1$ and $y = 0$ at $t = 1$, an author will appear at the next conference with probability $g(x, y)$ or skip it with the probability $1 - g(x, y)$. The author can terminate his/her membership in the community at each time step with the probability H . In order to estimate the values of parameters p , y_0 and H , we calculate the distribution of total number of attendances x , from the simulations and compare it to the empirical distribution using Kullback-Leibler Distance [30]. We perform the simulations for several different sets of parameter (y_0, H, p) to determine which combination of parameter values makes the model optimally close to the empirical data. For each parameter set the results are averaged across 100 simulations.

Supporting Information

S1 File. Supplementary Information: A theoretical model for the associative nature of conference participation. Proportion of conference participants g with x conference attendances who are going to attend the conference next year (**Figure A**). Proportion of conference participants ρ with n missed conferences after x -th conference attendance who are going to skip the conference next year, but will take part at some future conference from the observation period (**Figure B**). Pages on the web from which we downloaded conference data (**Table A**). Summary of the conference data. Columns 2 and 3 indicate for each conference the year in which data we have collected begin (Y_0) and end (Y_f). The total number of different participants at the conference during that period of time is given in column 4 (**Table B**). The number of participants at the conference per year (**Table C**). Log likelihood ratio \mathcal{R} and the π -value compare the fit to the power-law with the fit to the power-law with an exponential cutoff for the probability distribution of number of conferences at which each author appears (**Table D**). Log likelihood ratio \mathcal{R} and the π -value compare the fit to the power-law with the fit to the power-law with an exponential cutoff for the probability distribution of the number of successive participations at the conference (**Table E**). Log likelihood ratio \mathcal{R} and the π -value compare the fit to the

exponential with the fit to the power-law with an exponential cutoff for the probability distribution of the time lag between two consecutive conference participations (**Table F**). The optimal parameter values of the model for each conference (**Table G**). Stagnancy rate $1 - g(1, 0)$ at $t = 1$ for each conference and exponent α of power-law with an exponential cutoff distribution fit with the corresponding conference order (**Table H**).
(PDF)

Acknowledgments

Numerical simulations were run on the PARADOX supercomputing facility at the Scientific Computing Laboratory of the Institute of Physics Belgrade.

Author Contributions

Conceived and designed the experiments: JS AC TK MMD. Analyzed the data: JS. Wrote the paper: JS AC TK MMD. Collected the empirical data: JS.

References

1. Lazer D, Pentland AS, Adamic L, Aral S, Barabasi AL, Brewer D, et al. Life in the network: the coming age of computational social science. *Science*. 2009; 323(5915):721–723. PMID: [19197046](#)
2. Cioffi-Revilla C. Computational social science. *CompStat*. 2010; 2(3):259–271.
3. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Rev Mod Phys*. 2009; 81:591–646. doi: [10.1103/RevModPhys.81.591](#)
4. Sen P, Chakrabarti BK. *Sociophysics: An Introduction*. Oxford University Press, Oxford; 2014.
5. Schweitzer F. Scientific networks and success in science. *EPJ Data Science*. 2014; 3:35. doi: [10.1140/epjds/s13688-014-0035-8](#)
6. Radicchi F, Fortunato S, Castellano C. Universality of citation distributions: Toward an objective measure of scientific impact. *Proc Natl Acad Sci USA*. 2008; 105(45):17268–17272. doi: [10.1073/pnas.0806977105](#) PMID: [18978030](#)
7. Radicchi F, Fortunato S, Markines B, Vespignani A. Diffusion of scientific credits and the ranking of scientists. *Phys Rev E*. 2009; 80(5):056103. doi: [10.1103/PhysRevE.80.056103](#)
8. Wang D, Song C, Barabási AL. Quantifying long-term scientific impact. *Science*. 2013; 342(6154):127–132. doi: [10.1126/science.1237825](#) PMID: [24092745](#)
9. Sinatra R, Wang D, Deville P, Song C, Barabasi AL. Scientific impact: the story of your big hit. In: *APS Meeting Abstracts*; 2014. p. 17004.
10. Petersen AM, Riccaboni M, Stanley HE, Pammolli F. Persistence and uncertainty in the academic career. *Proc Natl Acad Sci USA*. 2012; 109(14):5213–5218. doi: [10.1073/pnas.1121429109](#) PMID: [22431620](#)
11. Azoulay P, Graff Zivin JS, Wang J. Superstar Extinction. *QJ of Econ*. 2010; 125(2):549–589. doi: [10.1162/qjec.2010.125.2.549](#)
12. Petersen AM, Jung WS, Yang JS, Stanley HE. Quantitative and empirical demonstration of the Matthew effect in a study of career longevity. *Proc Natl Acad Sci USA*. 2011; 108(1):18–23. doi: [10.1073/pnas.1016733108](#) PMID: [21173276](#)
13. Petersen A, Penner O. Inequality and cumulative advantage in science careers: a case study of high-impact journals. *EPJ Data Science*. 2014; 3(1):24. doi: [10.1140/epjds/s13688-014-0024-y](#)
14. Deville P, Wang D, Sinatra R, Song C, Blondel VD, Barabási AL. Career on the Move: Geography, Stratification, and Scientific Impact. *Sci Rep*. 2014; 4:4770. doi: [10.1038/srep04770](#) PMID: [24759743](#)
15. Borghans L, Romans M, Sauermann J. What makes a good conference? Analysing the preferences of labour economists. *Labour Economics*. 2010; 17(5):868–874.
16. Mair J, Thompson K. The UK association conference attendance decision-making process. *Tourism Management*. 2009; 30(3):400–409.
17. Witt SF, Sykes AM, Dartus M. Forecasting international conference attendance. *Tourism Management*. 1995; 16(8):559–570.
18. Garfield E. Making contacts at conferences. *The Scientist*. 1988; 2:294–295.

19. Garfield E. Making contacts at conferences—a problem for the young scientist. *Essays of an Information Scientist*. 1977; 3:668–672.
20. Van Dijk J, Maier G. ERSAs Conference participation: does location matter? *Papers in Regional Science*. 2006; 85(4):483–504. doi: [10.1111/j.1435-5957.2006.00102.x](https://doi.org/10.1111/j.1435-5957.2006.00102.x)
21. Majumdar SN. Persistence in nonequilibrium systems. *Current Science*. 1999; 77(3):370–375.
22. Drinea E, Frieze A, Mitzenmacher M. Balls and bins models with feedback. In: *Proceedings of the Thirteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA '02; 2002. p. 308–315.
23. Pemantle R. A survey of random processes with reinforcement. *Probab Surveys*. 2007; 4:1–79. doi: [10.1214/07-PS094](https://doi.org/10.1214/07-PS094)
24. Oliveira RI. The onset of dominance in balls-in-bins processes with feedback. *Random Structures & Algorithms*. 2009; 34(4):454–477. doi: [10.1002/rsa.20261](https://doi.org/10.1002/rsa.20261)
25. Johnson NL, Kotz S. *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*. Wiley; 1977.
26. Friedkin NE. Social cohesion. *Annu Rev Sociol*. 2004; 30:409–425. doi: [10.1146/annurev.soc.30.012703.110625](https://doi.org/10.1146/annurev.soc.30.012703.110625)
27. Sessions LF. How offline gatherings affect online communities. *Information, Communication & Society*. 2010; 13(3):375–395. doi: [10.1080/13691180903468954](https://doi.org/10.1080/13691180903468954)
28. Wu J, Ding XH. Author name disambiguation in scientific collaboration and mobility cases. *Scientometrics*. 2013; 96(3):683–697. doi: [10.1007/s11192-013-0978-8](https://doi.org/10.1007/s11192-013-0978-8)
29. Clauset A, Shalizi CR, Newman MEJ. Power-law distributions in empirical data. *SIAM Review*. 2009; 51(4):661–703. doi: [10.1137/070710111](https://doi.org/10.1137/070710111)
30. Burnham KP, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer Science & Business Media; 2002.

RESEARCH ARTICLE

Topology of Innovation Spaces in the Knowledge Networks Emerging through Questions-And-Answers

Miroslav Andjelković¹, Bosiljka Tadić^{2*}, Marija Mitrović Dankulov³, Milan Rajković¹, Roderick Melnik^{4,5}

1 Institute of Nuclear Sciences, Vinča, University of Belgrade, Belgrade, Serbia, **2** Department of Theoretical Physics, Jožef Stefan Institute, Ljubljana, Slovenia, **3** Scientific Computing Laboratory, Institute of Physics Belgrade, University of Belgrade, Zemun-Belgrade, Serbia, **4** MS2Discovery Interdisciplinary Research Institute, M²NeT Laboratory and Department of Mathematics, Wilfrid Laurier University, Waterloo, ON, Canada, **5** BCAM—Basque Center for Applied Mathematics, E48009 Bilbao, Basque Country—Spain

* bosiljka.tadic@ijs.si



OPEN ACCESS

Citation: Andjelković M, Tadić B, Mitrović Dankulov M, Rajković M, Melnik R (2016) Topology of Innovation Spaces in the Knowledge Networks Emerging through Questions-And-Answers. PLoS ONE 11(5): e0154655. doi:10.1371/journal.pone.0154655

Editor: Matjaz Perc, University of Maribor, SLOVENIA

Received: February 16, 2016

Accepted: April 15, 2016

Published: May 12, 2016

Copyright: © 2016 Andjelković et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The complete data are available on Stack Exchange site Mathematics <http://math.stackexchange.com/>. The exact data set which is used in this study is deposited into Figshare: https://figshare.com/articles/MathQuestions_data/3153145/MathQuestions.data.

Funding: This work was supported by Ministry of Education, Science, and Technological Development of the Republic of Serbia (<http://www.mpn.gov.rs/>), under the projects ON174014, ON171017; Research Agency of the Republic of Slovenia (<https://www.ars.gov.si/en/agencija/>), under the Program P1-0044; and

Abstract

The communication processes of knowledge creation represent a particular class of human dynamics where the expertise of individuals plays a substantial role, thus offering a unique possibility to study the structure of knowledge networks from online data. Here, we use the empirical evidence from questions-and-answers in mathematics to analyse the emergence of the network of knowledge contents (or tags) as the individual experts use them in the process. After removing extra edges from the network-associated graph, we apply the methods of algebraic topology of graphs to examine the structure of higher-order combinatorial spaces in networks for four consecutive time intervals. We find that the ranking distributions of the suitably scaled topological dimensions of nodes fall into a unique curve for all time intervals and filtering levels, suggesting a robust architecture of knowledge networks. Moreover, these networks preserve the logical structure of knowledge within emergent communities of nodes, labeled according to a standard mathematical classification scheme. Further, we investigate the appearance of new contents over time and their innovative combinations, which expand the knowledge network. In each network, we identify an innovation channel as a subgraph of triangles and larger simplices to which new tags attach. Our results show that the increasing topological complexity of the innovation channels contributes to network's architecture over different time periods, and is consistent with temporal correlations of the occurrence of new tags. The methodology applies to a wide class of data with the suitable temporal resolution and clearly identified knowledge-content units.

Introduction

The knowledge creation through online social interactions represents an emerging area of increased interest both for technological advances and the society [1] where the collective

Natural Sciences and Engineering Research Council of Canada, (<http://www.nserc-crsng.gc.ca/>), under the project code 213904.

Competing Interests: The authors have declared that no competing interests exist.

knowledge is recognised as a social value [2–4]. Recently studied examples include the knowledge accumulation in systems with direct questions-and-answers [5], crowdsourcing scientific knowledge production [6, 7] and scientific discovery games [8]. Similar phenomena can be observed in business/economics-associated online social networking [9–11]. On the other hand, the study of the collective knowledge creation opens new topics of research interests. In particular, it provides ground to examine a novel type of collective dynamics in social systems in which each actor possesses certain limited expertise. In the course of the collaborative social efforts to solve a problem, such as communications through questions-and-answers that we consider here, the tacit knowledge and the expertise of individual actors are externalised and dynamically shared with other participants who take part in the process. When a systematic tagging applies to the shared cognitive contents, the process leads to an explicit knowledge [3] as the output value (the network of knowledge contents), from which others can learn. Furthermore, the dynamics underlying knowledge creation exemplifies multi-scale phenomena related to the cognitive recognition, which may occur in a wider class of systems, social, biological and physical [17].

By the nature of the underlying stochastic processes, the knowledge networks that emerge through the collaborative social endeavours necessarily reflect the expertise and the activity patterns of the involved participants. Furthermore, these networks tend to capture the logical relationship among the used cognitive contents as it resides in the mind of each participating individual. In this regard, these networks substantially differ from the commonly studied knowledge networks, which are produced in ontological initiatives [12–14] such as those from the online bibliographic data and Wikipedia, or the mapping citation relationships between journal articles [15], to name a few. Also, the stochastic process of knowledge creation through questions and answers are different from the spreading dynamics of scientific memes, whose inheritance patterns are identified in citation networks [16].

In recent work [5], we have shown that the knowledge creation by questions-and-answers involve two-scale dynamics, in which the constitutive social and cognitive elements (individual experts or actors and the knowledge contents that they use) interact and influence each other on the original scale. This complex system evolves in a self-organised manner leading to the emergence of socio-technological structures where the involved actors share the accumulated knowledge. These structures are visualised as communities on the related bipartite network of actors and their artefacts [5]. Furthermore, the advance of innovation in this process, which builds on the expertise of the involved participants, leads to the expansion of the knowledge space by adding new cognitive contents. The central question for the research and applications of the collective knowledge creation is how these stochastic processes work and potentially can be controlled to converge towards the desired outcome. Furthermore, what is the structure of the emergent knowledge that can be used by others?

A part of the answer relies on the structure of the networks, co-evolving with the knowledge-sharing processes among the actors possessing the required expertise. In [5] the empirical data from the Stack Exchange site Mathematics (<http://math.stackexchange.com/>) were downloaded and analysed, as a prototypical example. The sequence of events in the process of questions-and-answers (Q&A) suitably maps onto a growing bipartite network of actors, as one partition, and their questions and answers, as another partition. The emergent communities on these networks have been identified, consisting of the involved actors and the connected questions-and-answers. As a rule, in each community a dominant actor is found, representing an active user with a broad expertise. The knowledge elements of each question are specified according to the standard mathematical classification scheme by one to five tags (for instance, “functional analysis”, “general topology”, “differential geometry”, “abstract algebra”, “algebraic number theory”). Consequently, the expertise of the actor can be specified as a combination of

tags that the actor had frequently used. Assuming that a minimal matching applies among the actor's expertise and the contents of the answered question, and using theoretical modelling based on the empirical data, it was shown [5] that the emergent communities and the knowledge that they share strongly depend on the population of the involved experts and their activity patterns.

In this work, using the same empirical dataset, our focus is on the networks of cognitive elements (tags) that emerge in these processes with questions-and-answers. Different from the aforementioned bipartite networks, these emergent knowledge networks contain subelements of both partitions, namely, knowledge contents of questions as well as a measure of the users' expertise. Such networks, supported by the current information and computer technology (ICT) systems, embody the collective knowledge that emerges via the cooperative social efforts and can be used by others to learn. Moreover, the relevance and speed of knowledge acquisition from these networks may be more efficient than from the networks generated through wide-scale ontological plans and efforts. We apply the techniques of algebraic topology of graphs [18–22] to investigate higher-order structures that characterise the connection complexity between knowledge elements in the emergent networks. Specifically, we aim to determine

- the metrics to quantify the higher-order combinatorial structures which contain the logical units of knowledge as the actors use them in communication;
- the role of innovative contents brought over time by the experts in building the network architecture.

In addition to the standard graph-theoretic metrics and community detection in the emergent networks of knowledge units, we describe their hierarchical organisation using several algebraic topology measures. Further, we identify the appearance of new tags over time and investigate the subgraphs (innovation channels) where these new cognitive elements attach to the existing network. By tracking topology measures over the consecutive time periods for the innovation channel together with the topology of the entire network, we quantify the impact of the new-added contents. Our main findings indicate that the networks of cognitive elements map to a nontrivial hierarchical architecture which contains aggregates of high-order cliques. The increasing structural complexity of these networks over time, owing to the innovation expansion, is consistent with the logical structure of knowledge that they contain and temporal correlations in the appearance of new cognitive contents.

In the following, the networks of tags are built from the empirical data for four successive one-year periods. At the initial stage, the networks are filtered to remove redundant links. At the next stage, network measures are obtained at the graph level, and the community structure is determined. At the final stage, the algebraic topology analysis of these networks for different periods and filtering levels is performed. The analysis is focused on the subgraphs, which are related to the appearance of new tags, representing the innovation channels of these networks.

Emergence of the tags networks

The Q&A process and structure of the empirical data

In this work, we have constructed knowledge networks from the empirical data, which are collected and described in Ref. [5]. In the data, the knowledge contents are mathematical tags used in the communications on Q&A system *Mathematics Stack Exchange*. In particular, the content of each question is specified (tagged) by one or more (maximum five) tags according to the standard mathematical classification scheme. While in Ref. [5] we investigated the role of expertise in the social process taking part on the co-evolving bipartite network of users-and-

questions, here we focus on the network of tags as the elementary units of knowledge that are used by the actors in this process. With the help of the agent-directed modeling, in Ref. [5] we have demonstrated that the considered empirical process obeys the fundamental assumption of knowledge creation, i.e., that at least minimal matching between the contents of the question and the expertise of answering actor occurred in each event. Therefore, the emergent network of tags reflects the way in which these knowledge units are used in the process and, indirectly, the expertise of the social community. Moreover, the architecture of the emergent network of tags is expected to mirror the logical structure of knowledge, as it is presented by the experts involved in the knowledge-creation process.

To be consistent with the previous studies and the associated analysis of Ref. [5], we use the same dataset that was downloaded on May 5, 2014, from <https://archive.org/details/stackexchange> and contains all user-contributed contents on Mathematics since the establishment of the site, July 2010, until the end of April 2014. Specifically, the considered dataset contains 269818 questions, posted and answered by 77895 users, 400511 answers, and 1265445 comments. For the present analysis, from the available high-resolution data we use the information about questions, i.e., ID of each question, its content as a list of tags, and time stamp. The tags and their combinations define the knowledge landscape whose size is not constant but increases with time and the number of posted questions. In this way, the innovation increases as the key feature of the collective knowledge creation [5]. By investigating the network of tags, here we examine how the knowledge creation can be expressed by the topological complexity of the expanding knowledge landscape.

Mapping data to networks of tags is performed within four consecutive periods; a period is one-year long. First, the questions that are posted within the considered year period are selected, and a unique set of tags that are involved in these questions is formed. Each tag represents a node of the tags network. Two tags (i, j) are linked by multiple connections w_{ij} , where the link multiplicity $w_{ij} = 0, 1, 2, \dots$ represents the number of common questions in which the considered pair of tags appeared in the selected dataset. The resulting networks are termed tagNetY-k, where $k = 1, 2, 3, 4$ indicates the considered year period.

Graph measures of tags networks without redundant connections

The raw networks of tags contain a large number of redundant connections leading to a large-density graph, cf. an example in Fig 1. To move forward, we first apply an advanced procedure to eliminate the potentially redundant links.

Filtering redundant connections in a network of tags is motivated by the following facts. In the data, the number of tags is between 500 and 1000 while the number of posted questions per year are between 15 and 120 thousand, which results in a quite dense network of tags. On the other hand, a broad distribution of the tags frequencies [5] suggests that a relatively small number of tags occurs quite frequently. Among the most frequent tags are “homework”, “proof-writing”, “reference-request”, and “terminology”, which are not related to any particular field of Mathematics but rather determine the type of question asked. For this reason, these tags can occur in many different combinations of tags, thus increasing the network’s density. Here, we apply an algorithm to decrease the network’s density by identifying the edges that do not incur as a result of a random process. For this purpose, the weighted network is considered as a multigraph where the weight w_{ij} represents a multiplicity of links between the pair of nodes (i, j) . We apply the filtering technique described in Ref. [23]; it utilizes a random configurational model for weighted graphs that preserves the total weight of the realised links, $W = \sum_k s_k$, as well as the node’s strength $s_k = \sum_j w_{ij}$ on average. To avoid the influence of the filtering on higher structures, we apply the algorithm to each link independently.

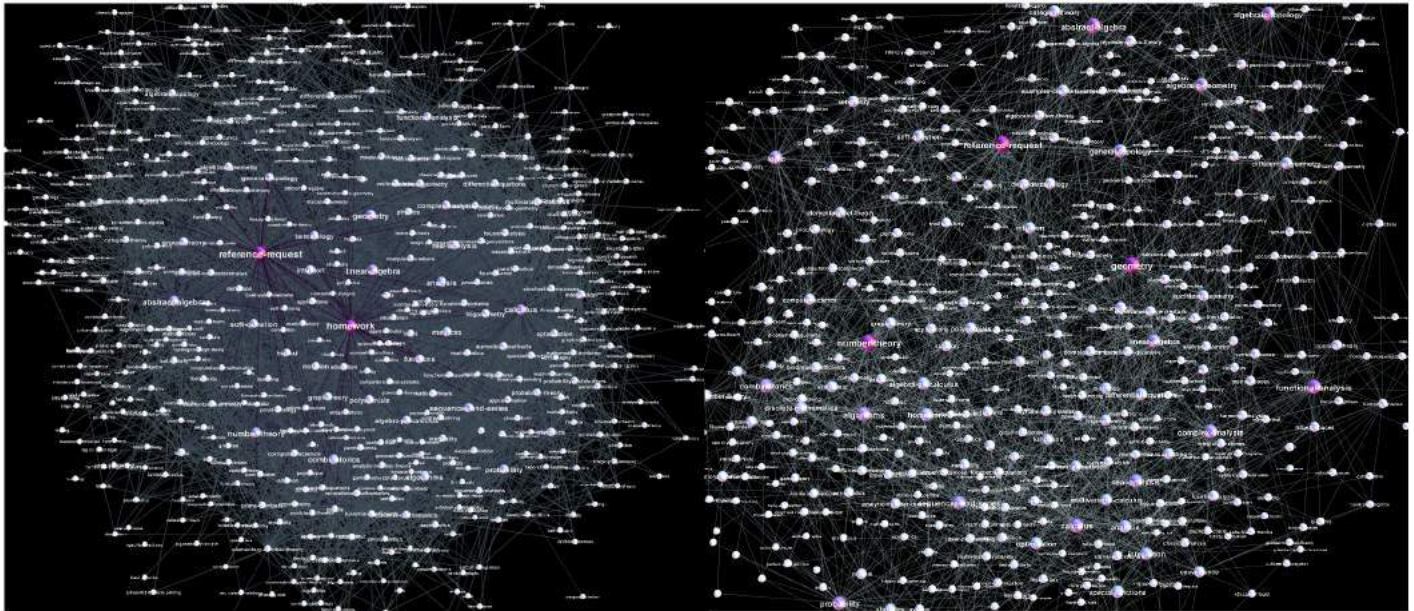


Fig 1. The network tagNetY-1: a close-up of unfiltered network near some large nodes (left) and the whole network filtered at confidence level $p = 0.1$ (right).

doi:10.1371/journal.pone.0154655.g001

A pair of nodes (i, j) is selected proportionally to their strengths s_i and s_j . In the considered network, the selected pair is connected by the weighted link of the multiplicity w_{ij} . In the random configurational model, the occurrence of a link with multiplicity m between the selected pair of nodes is given by the conditional probability

$$P_{ij}(m|s_i, s_j, W) = \binom{W}{m} \left(\frac{s_i s_j}{2W^2}\right)^m \left(1 - \frac{s_i s_j}{2W^2}\right)^{W-m}. \quad (1)$$

Then the probability that the realised weight w_{ij} of the link (i, j) occurred by chance (p -value) according to the marginal distribution given by Eq (1) is computed as [23]

$$P_r(w_{ij}) = \sum_{m \geq w_{ij}} P_{ij}(m|s_i, s_j, W). \quad (2)$$

The links for which the probability $P_r(w_{ij})$ appears to be larger than a preset confidence level p are removed. The remaining edges, which satisfy the condition $P_r(w_{ij}) \leq p$, represent the filtered network with the specified confidence level. Here we examine the structure of the filtered networks obtained for several values of the parameter, $p \in \{0.1, 0.05, 0.01\}$. As an example, the right panel in Fig 1 shows the first year network after the filtering procedure with the confidence level $p = 0.1$.

The networks of tags for different periods and filtered at various confidence levels are analysed by algebraic topology techniques, as presented in the following Sections. In this regard, we turn the weighted networks into binary graphs, which retain all important topological features of the weighted graphs while making the computation less demanding. Here, we first show that the filtering process leads to a reduced-density graph but preserves the relevant (nonrandom) connections. Specifically, the thematically connected groups of nodes (cf. labels of nodes in Figs 1 and 2) appear to form distinct communities on the network. In these

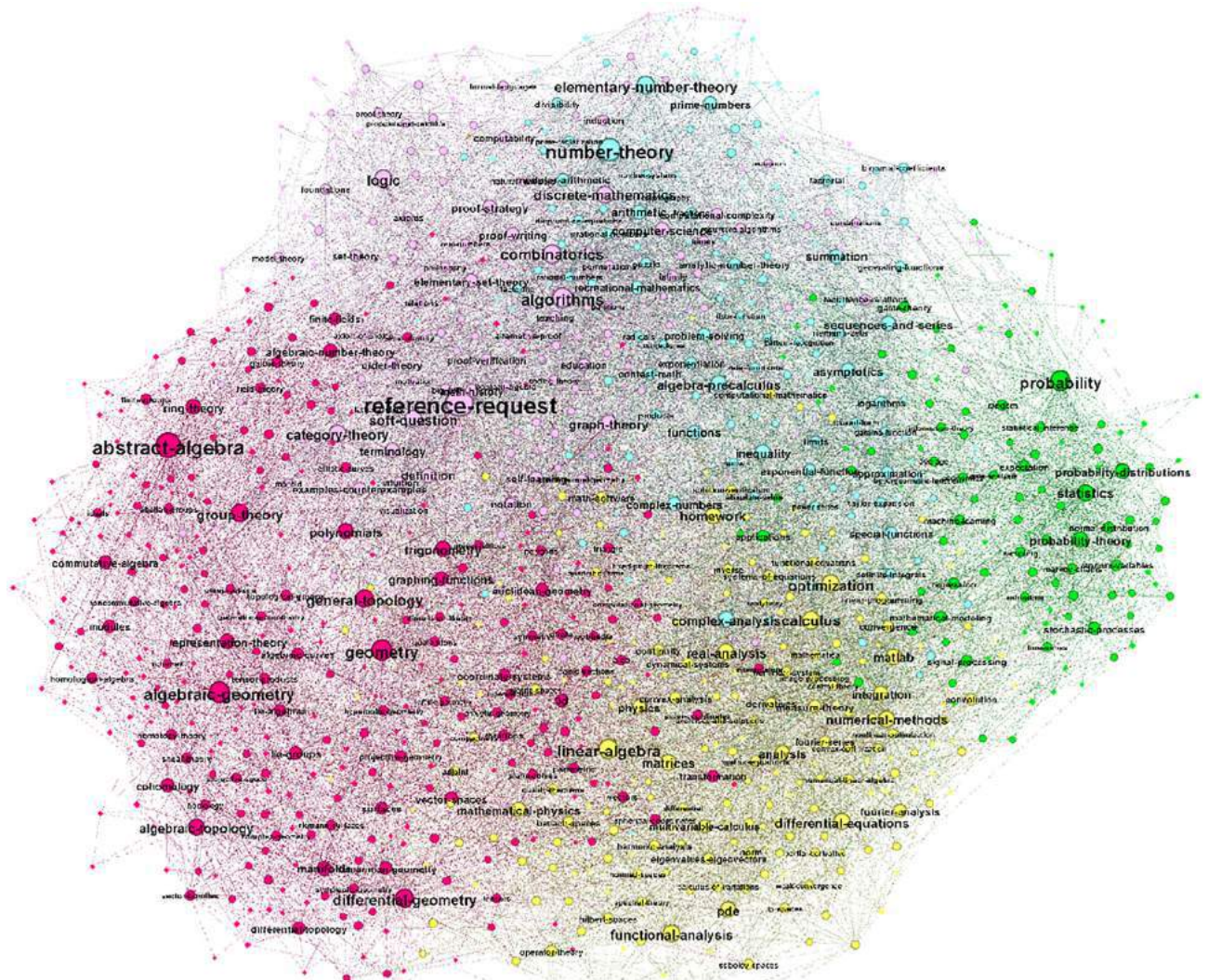


Fig 2. The community structure of the network of tags for the fourth period, which is filtered at $p = 0.1$. In each community, the mutually connected cognitive contents (mathematical tags) are indicated by the nodes' labels.

doi:10.1371/journal.pone.0154655.g002

networks, mostly non-overlapping communities occur. Consequently, they are suitably identified by methods based on the optimisation of the modularity [24–26]. A module is recognised as a densely connected group of nodes that are sparsely connected to nodes in other groups [27]. For a better comparison of different networks, the communities are systematically determined at the same resolution parameter (standard resolution 1.0 in Gephi, the open graph visualization platform <http://gephi.org>). This large-scale clustering of the knowledge networks appears systematically during the network growth. See also the structure of innovation channels studied in the following Section.

For comparison, in Table 1 we summarise the standard graph-theoretic measures [27] of the networks of tags for four consecutive periods and the confidence level $p = 0.1$. Note that the network of tags grows over years by the appearance of new tags, but also shrinks by the number of tags that appeared in the previous period and were not used in the current period.

Table 1. The graph-level measures for tags networks for four consecutive periods, filtered at confidence $\rho = 0.1$.

Net	N	$\langle k \rangle$	$\langle \ell \rangle$	d	Cc	ρ	M
TagNetY-1	582	10.07	3.02	6	0.365	0.018	0.439
TagNetY-2	702	14.45	2.86	5	0.365	0.021	0.441
TagNetY-3	856	20.09	2.72	5	0.351	0.023	0.436
TagNetY-4	1033	22.52	2.68	5	0.338	0.022	0.422

The number of nodes N , average degree $\langle k \rangle$, average path length $\langle \ell \rangle$, diameter d , clustering coefficient Cc, graph density $\rho = \frac{L}{N(N-1)}$, and modularity $M = \sum_i (e_{ii} - (\sum_j e_{ij})^2)$, where the summation runs over different communities.

doi:10.1371/journal.pone.0154655.t001

Topology of the tags networks

In addition to the standard graph-theoretic analysis, cf. Table 1, we apply techniques of algebraic topology to determine simplices and simplicial complexes, which describe higher order structures of these networks. Definitions and detailed explanation of topological quantities used in this presentation may be found in Ref. [19] and references within. The simplices are identified as maximal cliques of all orders, i.e. dimensions. Then the topological complexity of the simplicial complex constructed from the complex network is quantified by the number of cliques at each topological level (dimension) q , starting at $q = 0$ up to the $q_{max} - 1$. A clique at level $q = 0$ is an isolated node while $q = 1$ is a link, $q = 2$ is a triangle and so on up to the level $q_{max} - 1$ representing the largest clique found in the network.

Algebraic topology measures

We use the Bron-Kerbosh algorithm [21, 22] to determine cliques of all orders that are present in the studied network. The resulting matrix of maximal cliques (MC) thus contains information about the identity index of each clique as well as the identity index of each node that participates in that clique. Using rich information of the MC matrix, we can characterise the topological spaces around each node as well as the organisation of cliques in the entire network at each topological level. These goals are achieved by determining several node-related quantities [19] in addition to the commonly defined structure vectors of the network [18–20, 28].

In particular, the topology vector \mathbf{Q}^i is associated with the node i

$$\mathbf{Q}^i = \{Q_{q_{max}-1}^i, Q_{q_{max}-2}^i, \dots, Q_0^i\}, \tag{3}$$

where the components $Q_k^i, k = 0, 1, \dots, q_{max} - 1$, describe the number of k -dimensional cliques in which the node i participates. Then the influence of a node in the overall network architecture is quantified by *topological dimension* $dimQ^i$ of the node i , which is introduced in [19]; it is defined as the total number of all cliques in which the node i participates

$$dimQ^i = \sum_{q=1}^{q_{max}-1} Q_q^i. \tag{4}$$

To demonstrate the relevance of nodes, we compute the topological dimension of each node in the original and filtered network of tags for the first-year interval, which are shown in Fig 1. The components at each q level of the top 40 nodes (tags), ordered according to their topological dimension, are displayed by three-dimensional plots in Fig 3. As this figure shows, the applied elimination of the links reduces not only the node’s topological dimension but also changes the structures at q -levels where the considered node is present. Consequently, the

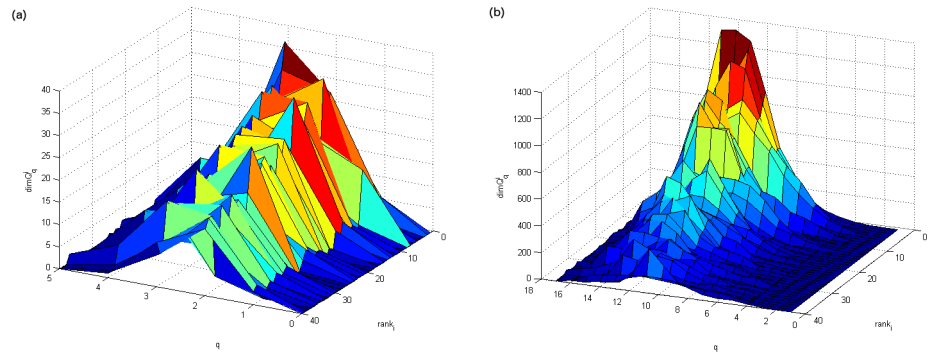


Fig 3. Components Q_q of the first 40 tags ranked by their topological $dim Q$ for the tagNetY-1 network filtered at $p = 0.1$ (a) and with no filtration (b).

doi:10.1371/journal.pone.0154655.g003

ranking order of a particular node can be changed (see the corresponding lists of nodes in Table 2), which is compatible with the altered importance of that node in the filtered network.

We further compare the role of individual nodes in the networks evolving over time, which are filtered at different confidence levels, i.e., $p = 0.1$, $p = 0.05$ and $p = 0.01$. We determine the topological dimensions of all nodes in the corresponding filtered networks for the four successive year-periods. The ranking distributions of the node’s topological dimensions are displayed in Fig 4a. This Figure shows that, first, nodes with a gradually higher topological dimension appear at later periods, suggesting that topological complexity of tags networks increases over years. Furthermore, within each year, the reduced confidence level p results in a simpler structure of the nodes’ neighbourhood (and possible shifts in the ranking order of nodes, as

Table 2. Names of the first twenty tags ordered according to their topological dimension in the network of tags before filtering and after filtering at the indicated confidence level p has been performed.

before filtering	$p = 0.1$	$p = 0.05$	$p = 0.01$
calculus	number theory	number theory	geometry
linear algebra	geometry	geometry	number theory
analysis	algebraic topology	functional analysis	calculus
homework	combinatorics	sequences and series	algorithms
reference request	abstract algebra	combinatorics	functional analysis
probability	functional analysis	algebraic topology	abstract algebra
sequences and series	algebra precalculus	abstract algebra	reference request
geometry	group theory	differential geometry	real analysis
functions	real analysis	calculus	algebra precalculus
real analysis	differential geometry	real analysis	logic
combinatorics	logic	probability	sequences and series
abstract algebra	sequences and series	algebraic geometry	probability
number theory	soft question	algebra precalculus	linear algebra
terminology	probability	algorithms	algebraic topology
complex analysis	integration	soft question	combinatorics
general topology	algorithms	logic	complex analysis
category theory	complex analysis	analysis	differential geometry
algebraic geometry	analysis	complex analysis	soft question
discrete mathematics	differential equations	integration	discrete mathematics
logic	calculus	differential equations	analysis

doi:10.1371/journal.pone.0154655.t002

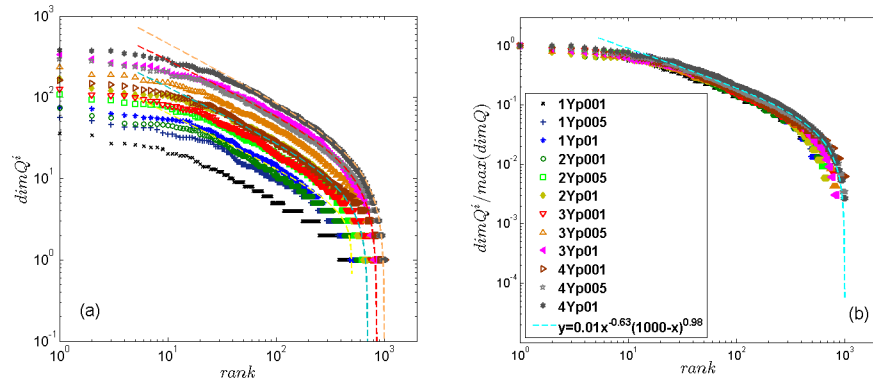


Fig 4. Ranking distributions of the topological dimension of tags $dimQ^i$ for all years and all p values (a) and scaled distribution $dimQ^i / \max(dimQ^i)$ of all data (b). The legend abbreviations: 1Yp001 indicates the first-year network filtered at the level $p = 0.01$, and so on. Fit lines are according to the discrete generalised beta function (5); in panel (a) the parameter $b = 0.67 \pm 0.03$ and c varies from 0.32 for 1Yp001 and 0.71 for 2Yp001 to 0.82 for 3Yp001 and 4Yp001, with error bars ± 0.03 .

doi:10.1371/journal.pone.0154655.g004

mentioned above). However, all networks exhibit a broad ranking distribution of the node’s topological dimension with a power-law section. The distributions are fitted by the discrete generalised beta function

$$f(x) = ax^{-b}(N + 1 - x)^c \tag{5}$$

with different parameters a , b and c . The robustness of the observed scaling feature is further confirmed by the scaling collapse of all curves to a master curve, shown in Fig 4b. The scale-invariant ranking, where the node’s topological dimension is scaled by the maximal dimension found in the corresponding network, suggests that the relative topological complexity of the tags networks is preserved over time and the degree of filtering.

Topological spaces in the filtered networks of tags

To characterise the structure of the topological levels $q = 0, 1, 2, \dots, K$ of the entire graph, we compute three commonly used structure vectors [18–20, 28, 29]. In particular, the first structure vector

$$\mathbf{Q} = \{Q_{q=K}, Q_{q=K-1}, \dots, Q_{q=1}, Q_{q=0}\} \tag{6}$$

has $K + 1$ components that describe the number of q -connected classes, where $K + 1 = q_{max}$ indicates the size of the maximal clique found in the graph. Furthermore, the components of the second structure vector

$$\mathbf{N}_s = \{n_{q=K}, n_{q=K-1}, \dots, n_{q=1}, n_{q=0}\} \tag{7}$$

designate the number of simplexes from the level q up to the top level. The third structure vector is often defined such that its q -level component

$$\hat{Q}_q = 1 - \frac{Q_q}{n_q} \tag{8}$$

determines how the simplices of higher order are connected at the level q . Fig 5 summarises the components of two structure vectors for the tags networks emerging over different periods and varied filtering level p .

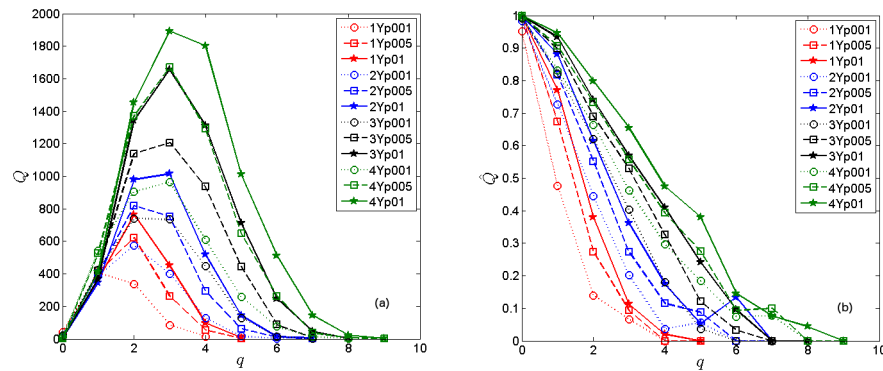


Fig 5. The components of (a) the first structure vector Q_q and (b) the third structure vector $\hat{Q}_q = 1 - n_q/Q_q$ plotted against the topology level q for each year period and three filtering levels $p = 0.1, 0.05, 0.01$. The legend abbreviations are explained in connection to Fig 4.

doi:10.1371/journal.pone.0154655.g005

By comparing the curves for different one-year periods but fixed filtering level, say $p = 0.1$, we observe that the network topological complexity increases over time. It manifests in the increased number of connectivity classes (components of the first structure vector) at all topological levels as well as the shift of the maximum from $q = 2$ (triangles), in the first year, to $q = 3$ (tetrahedra) and $q = 4$ (5-cliques), in the fourth year. At the same time, we observe that the number of topological levels increases as well as the connectivity among the cliques at each topology level, cf. the third structure vector in the Fig 5b.

On the other hand, by decreasing the filtering confidence level p , a more sparse network is obtained having a smaller number of topological levels and a reduced number of simplicial complexes. However, they proportionally preserve the above-described tendency of the enhanced complexity of combinatorial spaces over time. The corresponding curves for $p = 0.05$ and $p = 0.01$ are also shown for each year-period in Fig 5. According to the structure vectors in Fig 5, all filtered networks exhibit a systematic shift towards richer topology in later years. Once again, these results confirm the structural stability in Fig 4 of the emergent networks of tags, which complements the logical organisation of knowledge contents in the communities in these networks, demonstrated in Fig 2 and in the following Section.

Clustering of the innovative contents

Three aspects of innovation in the knowledge creation

The innovation growth [5, 30] is a crucial element of the process of knowledge creation. In the voluntary system, the innovation that comes from the expertise of the actors involved in the process was shown [5] to expand the knowledge space over time. To quantify the impact of innovation onto the architecture of the emerging knowledge networks, we consider the following three aspects of the innovation:

- the appearance of new tags due to the actor’s expertise;
- the occurrence of new combinations of tags expanding the knowledge space;
- the emergence of new combinatorial topological structures enriching the architecture of the knowledge network.

In the following, we discuss in detail these features of innovation.

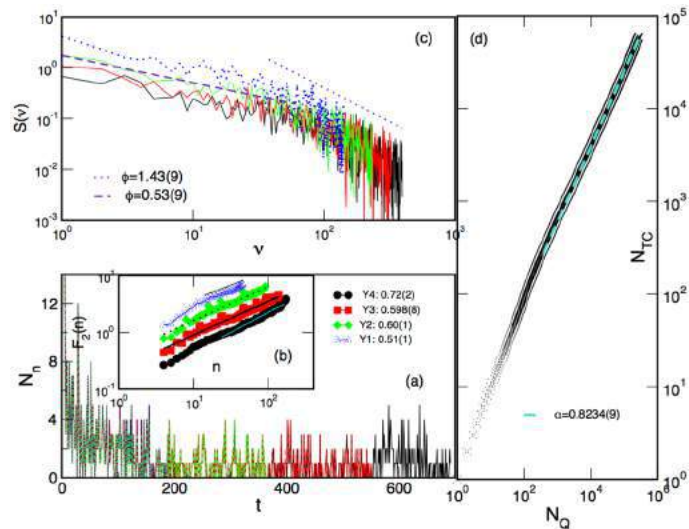


Fig 6. The temporal sequence of the appearance of new tags present in the networks for Year-1, Year-2, Year-3 and Year-4 periods (a). Temporal resolution is two days. The scaling of the standard fluctuation function (b) and the power spectrum (c) of these time series. Panel (d) displays increase in the number of new combinations of tags as a function of the number of questions over time.

doi:10.1371/journal.pone.0154655.g006

Fig 6a contains time sequence of the first appearance of tags that are present in the data of each one-year period. Naturally, the sequence for Year-1 is the shortest, while the sequence for Year-4 is the longest, since some tags that are present in Year-4 appeared in the earlier periods. The time series contains the number of new tags appearing in the sequence of two-day time intervals. The fractal analysis of these time series and their power spectrum, shown in Fig 6b and 6c suggest that the appearance of new tags is not random but exhibits long-range temporal correlations. Specifically, the plots in Fig 6b represent the fluctuation function $F_2(n)$ of the standard deviations of the integrated time series at the interval of length n . They reveal scaling regions (of different length for each time series) which permit determination of the Hurst exponent via $F_2(n) \sim n^H$. Values of the Hurst exponent H indicated in the legend suggest the fractal structure of the fluctuations. It appears that the fractality increases over time from nearly random time series with $H = 0.51 \pm 0.01$ in Year-1, to strongly persistent fluctuations with $H = 0.72 \pm 0.02$, in Year-4.

Similarly, power spectra of these time series in Fig 6c exhibit long-range correlations according to $S(v) \sim v^{-\phi}$ with two distinct exponents in high and low frequency regions. While the low-frequency feature is similar for all considered periods, the pronounced scaling in the high-frequency region gradually builds over years.

The number of unique combinations of tags was examined in the whole dataset and plotted against the number of posted questions in Fig 6d. The plot exhibits a power-law behaviour $N_{TC} \sim N_Q^\alpha$ in the range above 10^2 posted questions. It represents the Heaps' law which appears to be in agreement with the ranking distribution of frequencies of the unique combinations of tags, i.e., the Zipf's law, as discussed in [5]. The occurrence of Heaps' law is a manifestation of the innovation growth [5, 30] in the process of Q&A. The exponent $\alpha < 1$ indicates a sublinear growth of innovation with the number of posted questions. This dependence suggests that a fraction of displayed items brings new combinations of tags while the remaining questions use the already identified combinations.

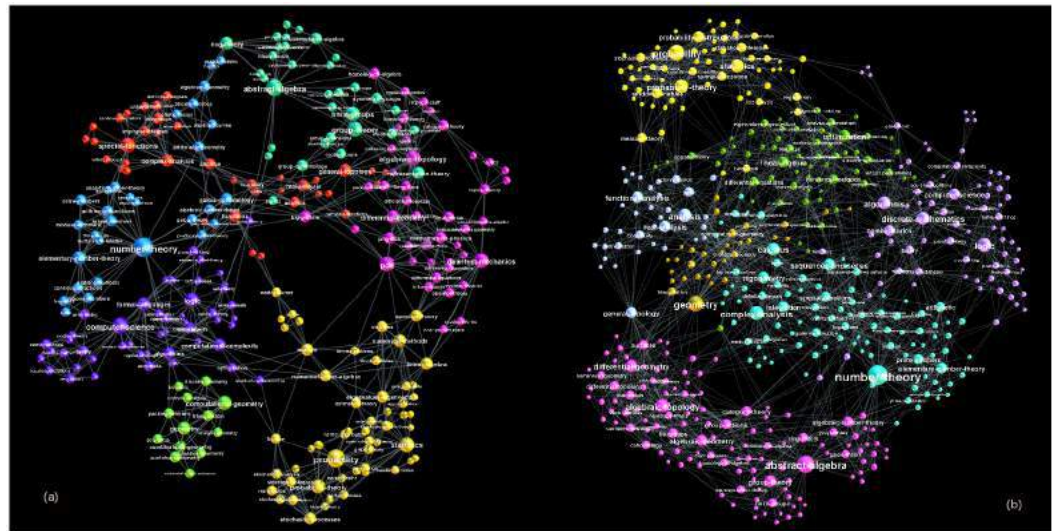


Fig 7. The structure of the innovation channel at the beginning of Year-2 (left) and Year-3 (right). New tags were added to the filtered tags network of the previous year, forming structures of higher dimension than a triangle. Communities of well-connected nodes show the logical grouping of mathematics subject categories, indicated by labels on nodes.

doi:10.1371/journal.pone.0154655.g007

The structure of innovation subgraphs

The appearance of new tags in the Q&A process leads to the expansion of the knowledge network. In particular, the network grows by the addition of new nodes (cf. Table 1), as well as by increasing its topological complexity measured by the presence of simplicial complexes of a high order. In the remaining part of this section, we investigate how the new tags attach to the existing nodes and affect the formation of higher order structures in the knowledge network. For this purpose, we first define an *innovation channel* as a subgraph related with the new tags appearing at the end of a considered one-year period. Specifically, the subgraph in the network (filtered at $p = 0.1$) contains newly added tags together with the tags to which they attach and form simplices larger than a single link (i.e., triangle or higher dimensional structure). The two plots in Fig 7 show the structure of the innovation channels at the beginning of Year-2 and Year-3 periods, respectively.

The innovation channels in Fig 7 grow over a one-year period; moreover, the innovative nodes stick with the rest of the network (previously existing nodes and links) making with them a tight structure that involves higher-order combinatorial spaces up to the largest clique. The community structure in the innovation subgraphs, which is demonstrated in Fig 7, reflects the thematic grouping of the entire knowledge network, as presented in Fig 2. For example, the newly added tag “cohomology” sticks to the group where we also find “algebraic topology”, “differential geometry”, “abstract algebra”, “complex geometry” and other thematically related tags, cf. the lower left community in Fig 7 right panel. On the other hand, the added tag “computational complexity” links to the community with “discrete mathematics”, “algorithms”, “logics”, “combinatorics”, “computer science” and others, cf. the rightmost community in the same Figure. Similarly, the node labels in all identified communities confirm their thematic closeness. Therefore, the expansion of the knowledge network by the addition of innovative contents systematically obeys the overall logical structure of (mathematical) knowledge. As mentioned earlier, the core of this feature of knowledge creation lies in the crucial role of the actor’s expertise in the process of meaningful cognitive-matching

actions. The logical structure of individual knowledge of each actor gets externalised during the process of Q&A.

According to the results in Fig 6, the appearance of innovative contents boosts the process of knowledge creation, leading to the observed temporal correlations, characteristic of collective dynamics. Analogously, here we show that the structure of innovation channels enriches the topological spaces of the knowledge network. In Figs 8 and 9 we summarise the topological measures of the innovation channels and compare them with the corresponding measures of the entire network. In addition to the structure vectors defined in Eqs (6)–(8), here we also consider the topological “response” function f_q to express the shifts in the topology at each level q in response to the changes in the network size. Formally, f_q is defined [20] as the number of simplices and shared faces at the level q .

Interestingly enough, the third structure vectors in Fig 8a and 8c show that the corresponding channels exhibit a better connectivity up to the level $q = 4$ of 5-clique than the background network. This feature of the innovation channels suggests the leading role of the innovative tags in the observed increase of the topological complexity of the network over years. This conclusion compares well with the number of connectivity classes at different topological levels, namely the first structure vectors in Fig 8b and 8d. The topology of the channel determines the most ubiquitous structure in the entire network, corresponding to the peak in the first structure vector. Furthermore, the increase of the topological complexity of the knowledge graphs over consecutive periods is illustrated by the topological “response” function f_q , which is shown in

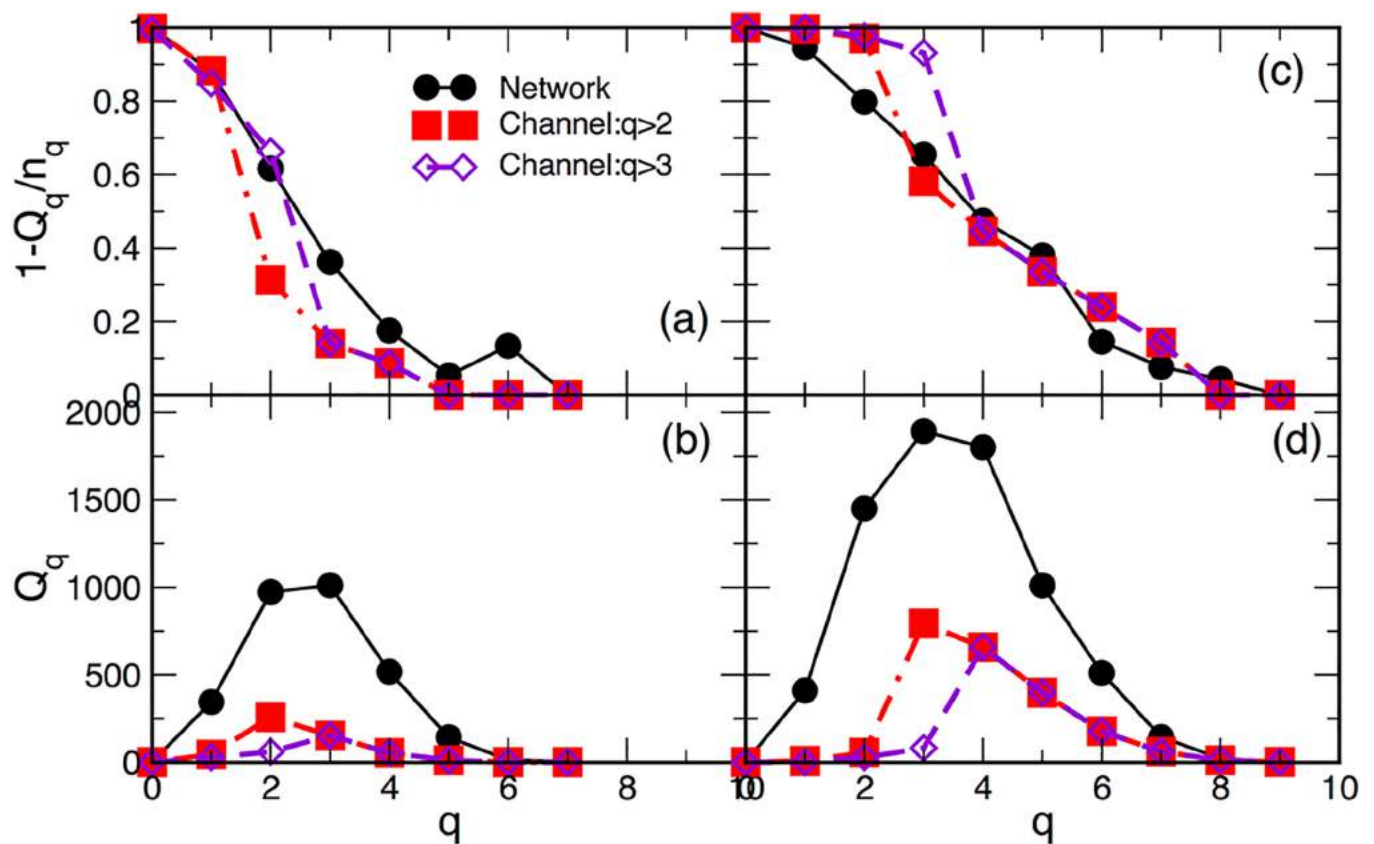


Fig 8. (a) and (c) The third structure vector and (b) and (d) the first structure vector for the networks of tags in Year-2 (left panels) and Year-4 (right panels) and the corresponding innovation channels above the level $q = 2$ and $q = 3$.

doi:10.1371/journal.pone.0154655.g008

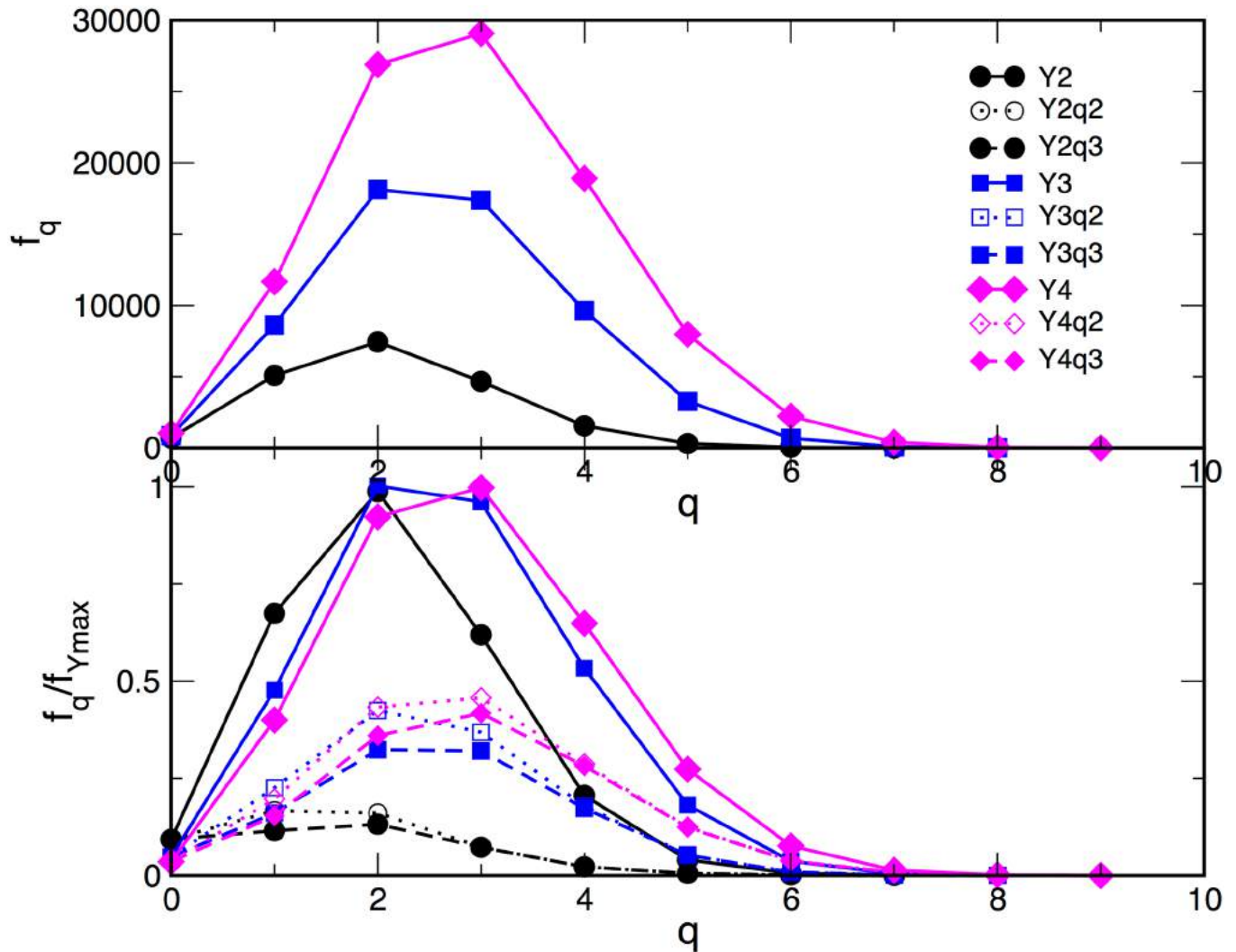


Fig 9. Response f_q plotted against the topology level q for networks of Year-2, Year-3, and Year-4 (top panel) and for the corresponding innovation channels scaled with the year maximum value (bottom panel).

doi:10.1371/journal.pone.0154655.g009

[Fig 9](#). It manifests in the increase of the number of topology levels, as well as the number of simplices and shared faces at each topology level. Also, the maximum of the function f_q shifts towards more complex structures, i.e., from triangles at Year-2 to tetrahedra in Year-4. As the plots in the lower panel of [Fig 9](#) show, these topological shifts in the networks of different periods are tightly reflected in the structure of the corresponding innovation channels.

Conclusions

Information processing underlines the evolution and structure of various social networks [31–33]. The creation of knowledge through questions-and-answers requires meaningful interactions with the actor’s expertise adjusted to the needs of other participants; consequently, it leads to the accumulation of the sound knowledge and the expansion of knowledge space [5]. In the studied example, we have demonstrated how the algebraic topology measures can characterise the connection complexity of the emergent knowledge networks. Using the data of questions-and-answers from the Stack Exchange system Mathematics, we have shown how the

network of mathematical tags, as constitutive elements of knowledge, appears and evolves with the actor–question–actor–answer interactions over time.

The connections among different tags reflect their use by the actors possessing the expertise, which (at least partially) overlaps with the contents of the considered question. The networks of tags are filtered by removing the extra edges which may have appeared by chance with a specified confidence level. We have applied the filtering at the level of (uncorrelated) edges to preserve the higher-order structures, which have been the focus of this study. Our results reveal that the process preserves the genuine structure of knowledge networks consisting of thematically connected tags communities. For example, five communities in [Fig 2](#) appear in the filtered network of tags in Year-4. Considering the higher-order topological spaces, the filtered networks of tags exhibit a robust structure. The hierarchy of nodes sorted out according to their suitably scaled topological dimension is represented by a unique curve, independent of the evolution time and the filtering level.

The appearance of new contents (tags) over time plays a significant role in the process of knowledge creation and the related networks. As it was shown in [\[5\]](#), the occurrence of new contents and new combinations of contents are chiefly related to the expertise of newly arriving users. Therefore, the introduced combinations of tags obey the logical structure as it is presented by the participating experts. The growing number of unique combinations leads to the advance of innovation [\[5\]](#), as also shown in [Fig 6d](#). Moreover, their appearance is conditioned by the cognitive-matching interactions and the user's activity patterns. These features of the social dynamics are manifested in the non-random (persistent) fluctuations and long-range temporal correlations, as demonstrated in [Fig 6a, 6b and 6c](#). Further, the performed algebraic topology analysis has revealed the role of these innovative contents in building the architecture of knowledge network. Specifically, we have found that:

- the newly appearing tags connect to the current network at all levels from a single link to the cliques of the highest order;
- the innovation channel is recognised as a subgraph containing simplices larger than or equal to a triangle in which at least one of the new tags occurred; its growth and the increased topological complexity over time provides the evolution pattern of the entire network;
- the growth of the innovation channel is consistent with enhanced fractal features and temporal correlations of the appearance of new contents over time; it systematically obeys the sensible connections of contents, as also demonstrated in [Fig 7](#).

The presented results reveal that the creation of new combinations of knowledge contents (or innovation) is compatible with the non-random correlations in the sequence of new contents and their linking to the knowledge network. Hence the innovation expansion, as a core of each knowledge-creation process, can be additionally quantified by the fractal features of time series of new tags as well as the algebraic topology measures of the network's innovation channel. Hidden beneath these quantifiers of the emergent knowledge networks is the dynamics of human actors and their expertise, which provides the logical structure of the collective knowledge. Our approach consists of the appropriate data filtering, fractal analysis of time series, and algebraic topology techniques applied to the emergent knowledge networks and their innovative channels. The methodology can be useful to the analysis of a wide class of networks where the actors and their artefacts, as well as the cognitive elements used in the process, are clearly identified. These may include, among others, networks created by science, engineering, business and economics communities based on online collaborations. Further, such examples may also include a collection of articles (e.g. journal articles) referring to each other, where their

logical units are marked. In some such situations, keywords, memes, and concepts can be identified by machine learning methods.

Author Contributions

Conceived and designed the experiments: BT MMD MR RM. Performed the experiments: MA BT MMD. Analyzed the data: MA BT MMD MR RM. Contributed reagents/materials/analysis tools: MMD MR. Wrote the paper: BT RM MR.

References

1. Bowker GC, Leight Star S, Turner W, Gasser L. Social science, technical systems, and cooperative work. Psychology Press, New York, 2014.
2. Carpendale JIM, Müller U, editors. Social Interactions and the Development of Knowledge. Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey, 2013.
3. Kimmerle J, Kress U, Held Ch. The interplay between individual and collective knowledge: technologies for organisational learning. Knowledge Management Research & Practice (2010) 8, 33–44. doi: [10.1057/kmrp.2009.36](https://doi.org/10.1057/kmrp.2009.36)
4. Kitchener RF. Piaget's Social Epistemology, in "Social Interactions and the Development of Knowledge", editors Carpendale J.I.M. and Müller U., Lawrence Erlbaum Associates, Inc. Mahwah, New Jersey, 2013, pp. 45–66.
5. Mitrović Dankulov M, Melnik R, Tadić B. The dynamics of meaningful social interactions and the emergence of collective knowledge. Scientific Reports (2015) 5, 12197. doi: [10.1038/srep12197](https://doi.org/10.1038/srep12197)
6. Boudreau K, Gaule P, Lakhani KR, Riedl Ch, Woolley A. From crowd to collaborators: initiating effort and catalyzing interactions among online creative workers. Harvard Business School, working paper, 2014.
7. Lakhani KR, von Hippel E. How open source software works: "free" user-to-user assistance. Res. Policy (2003) 32, 923–943. doi: [10.1016/S0048-7333\(02\)00095-1](https://doi.org/10.1016/S0048-7333(02)00095-1)
8. Fortino G, Galzarano S, Gravina R, Li WA. A framework for collaborative computing and multi-sensor data fusion in body sensor networks. Infor. Fusion (2015) 22, 50–70. doi: [10.1016/j.inffus.2014.03.005](https://doi.org/10.1016/j.inffus.2014.03.005)
9. Sloan S, Bodey K, Gyrd-Jones R. Knowledge sharing in online brand communities. Quantitative Market Research (2015) 18(3), 320–345. doi: [10.1108/QMR-11-2013-0078](https://doi.org/10.1108/QMR-11-2013-0078)
10. Seraj M. We create, we connect, we respect, therefore we are: intellectual, social, and cultural value in online communities. J. Interactive Marketing (2012) 26, 209–222. doi: [10.1016/j.intmar.2012.03.002](https://doi.org/10.1016/j.intmar.2012.03.002)
11. Blasco-Arcas L, Hernandez-Ortega BI, Jimenez-Martinez J. Collaborating online: the roles of interactivity nad personalization. The Service Industries J. (2014) 34(8), 677–698. doi: [10.1080/02642069.2014.886190](https://doi.org/10.1080/02642069.2014.886190)
12. Thessen AE, Sims Parr C. Knowledge extraction and semantic annotation of text from the encyclopedia of life. PLoS ONE (2014) 9, e0089550. doi: [10.1371/journal.pone.0089550](https://doi.org/10.1371/journal.pone.0089550)
13. Leibon G, Rockmore DN. Orienteering in knowledge spaces: The hyperbolic geometry of wikipedia mathematics. PLoS ONE (2012) 8, e67508. doi: [10.1371/journal.pone.0067508](https://doi.org/10.1371/journal.pone.0067508)
14. Uddin Sh, Khan A., Baur LA. A framework to explore the knowledge structure of multidisciplinary research fields. PLOS ONE (2015) 10, e0123537. doi: [10.1371/journal.pone.0123537](https://doi.org/10.1371/journal.pone.0123537) PMID: [25915521](https://pubmed.ncbi.nlm.nih.gov/25915521/)
15. Fortunato S, Radicchi F, Vespignani A. in book Citation Networks. Springer-Verlag Berlin Heidelberg, 2012.
16. Kuhn T, Perc M, Heilbing D. Inheritance patterns in citation networks reveal scientific memes. Phy. Rev. X (2014) 4, 041036.
17. Kotsireas I, Melnik RVN, West B, editors. Advances in Mathematical and Computational Methods: Addressing Modern Challenges of Science, Technology and Society. American Institute of Physics (2011) Vol. 1368.
18. Jonsson J. Simplicial Complexes of Graphs. Lecture Notes in Mathematics, Springer-Verlag, Berlin, 2008.
19. Andjelković M, Tadić B, Maletić S, Rajković M. Hierarchical sequencing of online social graphs. Physica A: Statistical Mechanics and its Applications (2015) 436, 582–595. doi: [10.1016/j.physa.2015.05.075](https://doi.org/10.1016/j.physa.2015.05.075)
20. Andjelković M, Gupte N, Tadić B. Hidden geometry of traffic jamming. Phys. Rev. E (2015) 91, 052817. doi: [10.1103/PhysRevE.91.052817](https://doi.org/10.1103/PhysRevE.91.052817)

21. Bron C, Kerbosch J. Finding all cliques of an undirected graph. *Comm. ACM* (1973) 16, 575–577. doi: [10.1145/362342.362367](https://doi.org/10.1145/362342.362367)
22. Bandelt HJ, Chepoi V. Metric graph theory and geometry: a survey, in Goodman J. E.; Pach J.; Pollack R., eds. “Surveys on discrete and computational geometry: Twenty years later”. *Contemporary Mathematics* (2008) 453, 49–86. doi: [10.1090/conm/453/08795](https://doi.org/10.1090/conm/453/08795)
23. Diananti M. Unwinding the hairball graph: pruning algorithms for weighted complex networks. *Phys. Rev. E* (2016) 93, 012304. doi: [10.1103/PhysRevE.93.012304](https://doi.org/10.1103/PhysRevE.93.012304)
24. Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* (2008) 2008(10), P10008. doi: [10.1088/1742-5468/2008/10/P10008](https://doi.org/10.1088/1742-5468/2008/10/P10008)
25. Lancichinetti A, Kivela M, Saramaki J, Fortunato S. Characterizing the community structure of complex networks. *PLoS ONE* (2010) 5(8), e11976. doi: [10.1371/journal.pone.0011976](https://doi.org/10.1371/journal.pone.0011976) PMID: [20711338](https://pubmed.ncbi.nlm.nih.gov/20711338/)
26. Lambiotte R, Delvenne J-C, Barahona M. Laplacian Dynamics and Multiscale Modular Structure in Networks. *IEEE Transactions on Network Science and Engineering* (2015) 1(2), 76–90.
27. Dorogovtsev S. *Lectures on Complex Networks*. Oxford University Press, Inc., New York, NY, USA, 2010.
28. Maletić S, Horak D, Rajković M. Cooperation, conflict and higher-order structures of social networks. *Advances in Complex Systems* (2012) 15(supp01), 1250055. doi: [10.1142/S0219525912500555](https://doi.org/10.1142/S0219525912500555)
29. Maletić S, Rajković M. Consensus formation on a simplicial complex of opinions. *Physica A: Statistical Mechanics and its Applications* (2014) 397, 111–120. doi: [10.1016/j.physa.2013.12.001](https://doi.org/10.1016/j.physa.2013.12.001)
30. Tria F, Loreto V, Servedio VDP, Strogatz SH. The dynamics of correlated novelties. *Scientific Reports* (2014) 4, 5890. doi: [10.1038/srep05890](https://doi.org/10.1038/srep05890) PMID: [25080941](https://pubmed.ncbi.nlm.nih.gov/25080941/)
31. Sloot MAP, Quax R. Information processing as a paradigm to model and simulate complex systems. *Journal of Computational Science* (2012) 3, 247–249. doi: [10.1016/j.jocs.2012.07.001](https://doi.org/10.1016/j.jocs.2012.07.001)
32. Tadić B, Gligorijević V, Mitrović M, Šuvakov M. Co-evolutionary mechanisms of emotional bursts in online social dynamics and networks. *Entropy* (2013) 15(12), 5084–5120. doi: [10.3390/e15125084](https://doi.org/10.3390/e15125084)
33. Estrada E, Gómez-Gardeñes J. Communicability reveals a transition to coordinated behavior in multiplex networks. *Phys. Rev. E* (2014) 89, 042819. doi: [10.1103/PhysRevE.89.042819](https://doi.org/10.1103/PhysRevE.89.042819)

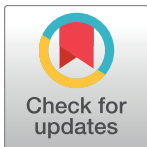
RESEARCH ARTICLE

Associative nature of event participation dynamics: A network theory approach

Jelena Smiljanić^{1,2*}, Marija Mitrović Dankulov¹

1 Scientific Computing Laboratory, Center for the Study of Complex Systems, Institute of Physics Belgrade, University of Belgrade, Pregrevica 118, 11080 Belgrade, Serbia, **2** School of Electrical Engineering, University of Belgrade, P.O. Box 35-54, 11120 Belgrade, Serbia

* jelenas@ipb.ac.rs



Abstract

The affiliation with various social groups can be a critical factor when it comes to quality of life of each individual, making such groups an essential element of every society. The group dynamics, longevity and effectiveness strongly depend on group's ability to attract new members and keep them engaged in group activities. It was shown that high heterogeneity of scientist's engagement in conference activities of the specific scientific community depends on the balance between the numbers of previous attendances and non-attendances and is directly related to scientist's association with that community. Here we show that the same holds for leisure groups of the Meetup website and further quantify individual members' association with the group. We examine how structure of personal social networks is evolving with the event attendance. Our results show that member's increasing engagement in the group activities is primarily associated with the strengthening of already existing ties and increase in the bonding social capital. We also show that Meetup social networks mostly grow through big events, while small events contribute to the groups cohesiveness.

OPEN ACCESS

Citation: Smiljanić J, Mitrović Dankulov M (2017) Associative nature of event participation dynamics: A network theory approach. PLoS ONE 12(2): e0171565. doi:10.1371/journal.pone.0171565

Editor: Matjaz Perc, University of Maribor, SLOVENIA

Received: November 15, 2016

Accepted: January 22, 2017

Published: February 6, 2017

Copyright: © 2017 Smiljanić, Mitrović Dankulov. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All data files are available at https://figshare.com/articles/Meetup_Datasets/2066904 (DOI: <https://dx.doi.org/10.6084/m9.figshare.2066904.v1>).

Funding: This work was supported by ON171017, The Ministry of Education, Science, and Technological Development of the Republic of Serbia (<http://www.mps.gov.rs/>: J.S. and M.M.D.); and 675121, The European Commission (http://ec.europa.eu/index_en.htm: M.M.D.).

Competing Interests: The authors have declared that no competing interests exist.

Introduction

One of the consequences of the rapid development of the Internet and growing presence of information communication technologies is that a large part of an individual's daily activities, both off and online, is regularly recorded and stored. This newly available data granted us a substantial insight into activities of a large number of individuals during long period of time and led to the development of new methods and tools, which enable better understanding of the dynamics of social groups [1]. The structure and features of social connections both have strong influence and depend on social processes such as cooperation [2, 3], diffusion of innovations [4, 5], and collective knowledge building [6]. Therefore, it is not surprising that the theory of complex networks has proven to be very successful in uncovering mechanisms governing the behavior of individuals and social groups [7, 8].

The human activity patterns, the structure of social networks, and the emergence of collective behavior in different online communities have been extensively studied in the last decade

[6, 9–15]. On the other hand, the dynamics of offline social groups, where the activities take place through offline meetings (events), have drawn relatively little attention, given their importance. Such offline groups, both professional and leisure ones, provide significant benefits and influence everyday lives of individuals, their broader communities, and the society in general: they offer social support to vulnerable individuals [16, 17], can be used for political campaigns and movements [18, 19], or can have an important role in career development [20]. As they have different purpose, they also vary in the structure of participants, dynamics of meetings, and organisation. Some groups, such as cancer support groups or scientific conference communities, are intended for a narrow circle of people while others, leisure groups for instance, bring together people of all professions and ages. In the pre-Internet era these groups have been, by their organisation and means of communication between their members, strictly offline, while today we are witnessing the appearance of a growing number of hybrid groups, which combine both online and offline communication [19]. Although inherently different, all these social groups have two main characteristics in common: they do not have formal organization, although their members follow certain written and non-written rules, and their membership is on a voluntary basis. Bearing this in mind, it is clear that the function, dynamics and longevity of such self-organized communities depend primarily on their ability to attract new and retain old active members. Understanding the reasons and uncovering key factors that influence members to remain active in the social group dynamics are thus important, especially having in mind their relevance for the broader social communities and the society.

The size of social groups and personal social networks, as well as their structure, have been extensively studied. The considerable body of evidence [21–24] suggests that the typical size of natural human communities is approximately 150, that both groups and personal social networks are highly structured, and consist of social layers characterized by different strengths of relationships. The relationships within each layer are characterized by a similar mean frequency of interaction and emotional closeness, both of which decrease rapidly as we move through network layer. These findings have been explained using the Social Brain Hypothesis, which relates the average size of species' personal network with the computational capacity of its brain. Here we confirm that these findings also hold for leisure groups where the frequency of interactions among members is constrained by the event dynamics. We also explore how the number of attended events is related to the size and layered structure of member's personal network.

Previous research on hybrid social groups and interplay between offline and online interactions has shown that offline meetings enhance attendees' engagement with online community and contributes to the creation of a bonding capital [25, 26]. In our previous work [20] it was shown that the participation patterns of scientist in a particular conference series are not random and that they exhibit a universal behaviour independent of conference subject, size or location. Using the empirical analysis and theoretical modeling we have shown that the conference attendance depends on the balance between the numbers of previous attendances and non-attendances and argued that this is driven by scientist's association with the conference community, i.e. with the number and strength of social ties with other members of the conference community. We also argued that similar behaviour can be expected in other social communities when it comes to members' participation patterns in organised group events. Here we provide empirical evidence supporting these claims and further investigate the relationship between the dynamics of participation of individuals in social group activities and the structure of their social networks.

The Meetup portal, whose group dynamics we study here, is an event-based social network. Meetup members use the online communication for the organization of offline gatherings. The online availability of event attendance lists and group membership lists enables us to

examine the event participation dynamics of Meetup groups and its influence on the structure of social networks between group members. The diversity of Meetup groups in terms of the type of activity and size allows us to further examine and confirm the universality of member's participation patterns. We note that previous papers using Meetup source of data have mostly focused on the event recommendation problem [27–31], structural properties of social networks, and relationships between event participants [31, 32] by disregarding evolutionary behaviour of Meetup groups.

In this paper, we examine the event-induced evolution of social networks for four large Meetup groups from different categories. Similarly to the case of conference participation, we study the probability distribution of a total number of meetup attendances and show that it also exhibits a truncated power law for all four groups. This finding suggests that the event participation dynamics of Meetup groups is characterized by a positive feedback mechanism, which is of social origin and is directly related to member's association with social community of the specific Meetup group. Using the theory of complex networks we examine in more detail the correlation between an individual's decisions to participate in an event and her association with other members of that Meetup group. Specifically, we track how member's connectedness with the community changes with the number of attendances by measuring change in the clustering coefficient and relation between the degree and the strength in an evolving weighted social network, where only statistically significant connections are considered. Our results indicate that greater involvement in group activities is more associated with the strengthening of existing than to creation of new ties. This is consistent with previous research on Meetup which has shown that repeated event attendance leads to an increase in bonding and a decrease in bridging social capital [25, 33]. Furthermore, in view of the fact that people interact and networks evolve through events, we examine how particular a event affects the network size and its structure. We investigate effects of event sizes and time ordering on social network organization by studying changes in the network topology, numbers of distinctive links and clustering, caused by the removal of a specific event. We find that large events facilitate new connections, while during the small events already acquainted members strengthen their interpersonal ties. Similar behavior was observed at the level of communities, where small communities are typically closed for new members, while contrary to this, changes in the membership in large communities are looked at favorably [34, 35].

This paper is organized as follows: we first study the distribution of the total number of participations in four Meetup groups from different categories. Next we introduce a filtered weighted social network to characterize significant social connections between members and discuss its structural properties. Specifically, we study how the local topological properties evolve with the growth of the number of participations in order to derive relationships between members' association with the group and their activity patterns. In order to analyze impact of a particular event on the network organization, we remove events using different strategies and study how this influences the social structure.

Results

Event participation patterns of Meetup groups

Meetup is an online social networking platform that enables people with a common interest to start a group with a purpose of arranging offline meetings (events, meetups) all over the world. The groups have various topics and are sorted into 33 different categories, such as careers, hobbies, socializing, health, etc. These groups are of various sizes, have different event dynamics, and hierarchical organisation. They also differ in the type of activity members engage, ranging from socializing events (parties and clubbing) to professional trainings (seminars and

Table 1. Summary of collected data for four selected Meetup groups. N_m is total number of group members, N_e is total number of organised events.

Meetup group	Acronym	Category	N_m	N_e
geamcIt	GEAM	Food & Drink	5377	3986
pittsburgh-free	PGHF	Socializing	4995	4617
techlifecolumbus	TECH	Tech	3217	3162
VegasHikers	LVHK	Outdoors & Adventure	6061	5096

doi:10.1371/journal.pone.0171565.t001

lectures). Common to all groups is the way they organize offline events: each member of the group gets an invitation to event to which they reply with yes or no, creating in that way a record of attendance for each event. We use this information to analyze event participation patterns and to study the evolution of the social network.

Here, in particular we analyze four large groups, each belonging to a different category and having more than three thousand organized events (see [Methods](#) and [Table 1](#)). We chose these four groups because of their convenience for statistical analysis, large number of members and organized events, and also for the fact that they are quite different concerning the type of activities and interests their members share. The *geamcIt* (GEAM) group is made of *foodie thrill-seekers* who mostly meet in restaurants and bars in order to try out new exciting foods and drinks, while people in the *VegasHiker* (LVHK) group are hikers who seek excitement through physical activity. The *Pittsburgh-free* (PGHF) is our third group which invites its members to free, or almost free, social events, and the fourth considered group *TechLife Columbus* (TECH), which is about social events and focuses on technology-related community networking, entrepreneurship, environmental sustainability, and professional development.

[Fig 1](#) demonstrates that the probability distributions of total attendance numbers of members in events for all four groups exhibits a truncated power law behavior (see [Fig A](#) and [Table A](#) in [S1 File](#), which show a comparison with the exponential and power law fit), with power law exponent larger than one. Similarly to the conference data [20], the exponential

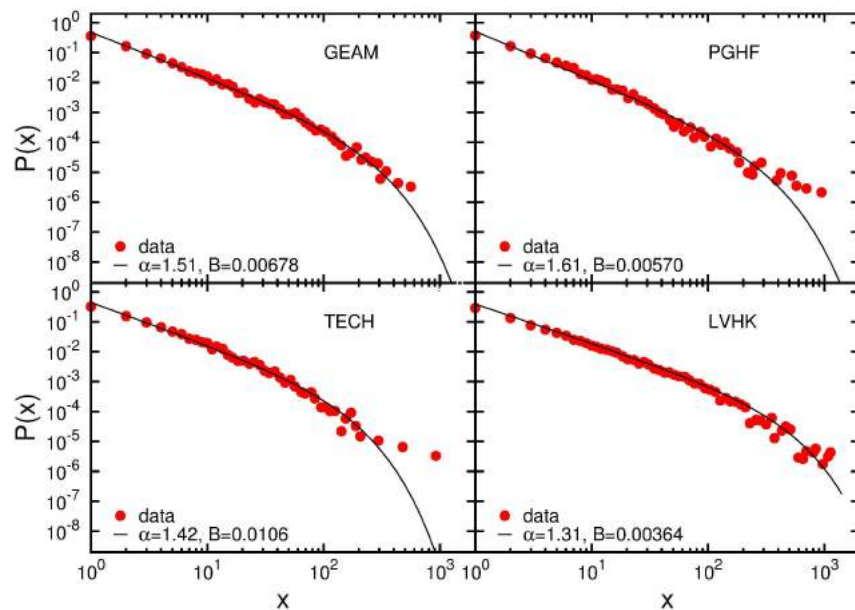


Fig 1. Total number of attended events. Probability distributions $P(x)$ of total number of participations x , for four Meetup groups. Solid line represents best fit to truncated power law distribution, $x^{-\alpha}e^{-Bx}$.

doi:10.1371/journal.pone.0171565.g001

cut-off is a finite size effect. Power law and truncated power law behavior of probability distributions can be observed for the number of and the time lag between two successive participations in group-organized events, Figs B and C in [S1 File](#). In fact, we find that similar participation patterns which differ in values of exponents) can be observed for all Meetup groups, regardless of their size, number of events or category. As in the case of the conference participation dynamics [20], this indicates that the probability to participate in the next event depends exclusively on the balance of numbers of previous participations and non-participations. We argued in [20] that the forces behind conference participation dynamics are of social origin, and it follows from [Fig 1](#) that the same can be argued for the case of the Meetup group participation dynamics. The more participations in group activities member has, the stronger and more numerous are her connections to the other group members, and thus her association with the community. We further explore this assumption by investigating the event-driven evolution of social networks of the four different Meetup groups.

Structure of social event-based network

We construct a social network between group members for each considered group, as a network of co-occurrence on the same event (see [Methods](#) for more details). By definition, these networks are weighted networks with link weights between two members equal to the number of events they participated together. These networks are very dense, as a direct result of the construction method, with broad distribution of link weights (see [Fig D](#) in [S1 File](#)). However, co-occurrence at the same event does not necessarily imply a relationship between two group members. For instance, a member of a group that attends many events, or big events, has a large number of acquaintances, and thus large number of social connections, which are not of equal importance regarding her association with the community. Similarly, two members that attend a large number of events can have relatively large number of co-occurrences, which can be the result of coincidences and not an indicator of their strong relationship. In order to filter out these less important connections we use a filtering technique based on the configuration model of bipartite networks [36, 37] (see [Methods](#)). By applying this technique to weighted networks we reduce their density and put more emphasis on the links that are less likely to be the result of coincidences. In this way we emphasize the links of higher weights without the removal of all links below certain threshold (see [Fig D](#) in [S1 File](#)), a standard procedure for network pruning. We explore the evolution of social networks of significant relationships between Meetup group members by studying how the local characteristics of the nodes (members) change with their growing number of participations in group activities.

Association with the community of a specific Meetup group can be quantitatively expressed through several local and global topological measures of weighted networks. Specifically, here we explore how the number of significant connections (member's degree) and their strength (member's strength), as well as how member's embeddedness in a group non-weighted and weighted clustering coefficient) are changing with the number of attended group events. [Fig 2](#) shows how average strength of a node depends on its degree in filtered networks of four selected Meetup groups. While member's degree equal to the number of member's significant social relationships, the strength measures how strongly she is connected to the rest of the group [38]. In all considered Meetup groups members with small and medium number of acquaintances ($q \leq 50$) have similar values of strengths and degrees, i.e., their association with the community is quantified by the number of people they know and not through the strength of their connections (see [Fig 2](#)). Having in mind that the average size of an event in these four groups is less than 20, we can conclude that majority of members with a degree less than 50 are the ones that attended only a few group meetups. A previous study [30] has found that the

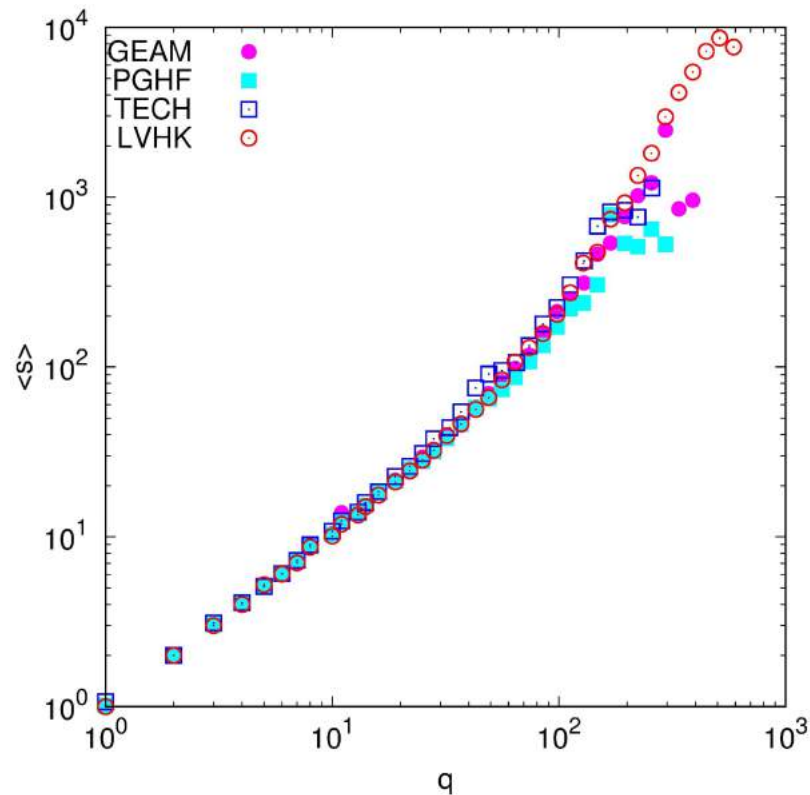


Fig 2. Node strength dependence on node degree. Dependence of average member's strength (s) on her degree q in social network of significant links for considered groups.

doi:10.1371/journal.pone.0171565.g002

probability for a member to attend a group event strongly depends on whether her friends will also attend. The non-linear relationship between the degree and the average strength for $q > 50$ shows that event participation of already engaged members (ones who already attended few meetings) is more linked to the strength of social relations than to their number. This means that at the beginning of their engagement in group activities, when the association is relatively small, the participation is conditioned by a number of members a person knows, while later, when the association becomes stronger, the intensity of relations with already known members becomes more important.

This finding is further supported if we consider the change of the average degree and strength with the number of participations. Fig 3 shows how the average member's degree and strength evolve with the number of participations in group's events. At the beginning, the degree and strength have the same value and grow at the same rate, but after only few participations the strength becomes larger than the degree, and starts to grow much faster for members of all four Meetup communities. After 100 attended events the average strength of a member is up to ten times larger than her degree (see Fig E in S1 File). This indicates that the event participation dynamics is mostly governed by the need of a member to maintain and strengthen her relationships with already known members of a community. As a matter of fact, our analysis of member's embeddedness in these social networks shows that members maintain strong relations with single members of the community, but also with small sub-groups of members. A comparison with randomized data (Figs E and F in S1 File) reveals that both the degree and strength grow slower with the number of events, and that their ratio is

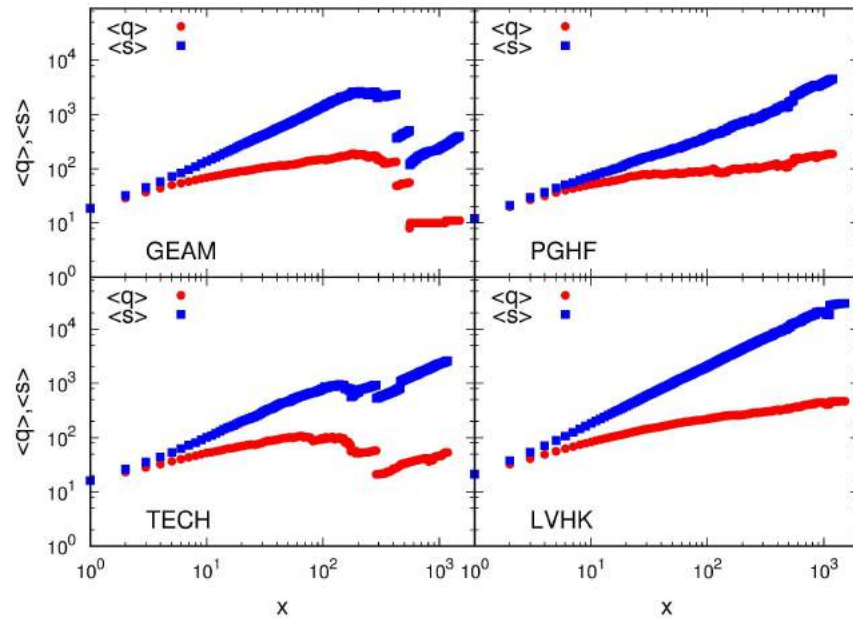


Fig 3. Event driven evolution of member's degree and strength. Dependence of member's average degree $\langle q \rangle$ and strength $\langle s \rangle$ on number of attended group events by member x for four considered Meetup groups.

doi:10.1371/journal.pone.0171565.g003

higher than in the original data. Relatively high value of the average clustering coefficient $\langle c_i \rangle$, shown in Fig 4 indicates that there is a high probability (more than 10% on average) that friends of a member also form significant relationships. The slow decay of $\langle c_i \rangle$ with the number of participations and the fact that it remains relatively large (above 0.2) even for participants with a thousand of attended meetups, Fig 4, show that personal networks of members have tendency to remain clustered, i.e., have relatively high number of closed triplets compared to random networks.

We now further examine the structure of these triplets and its change with the number of participations by calculating the averaged weighted clustering coefficient. The weighted clustering coefficient c_i^W measures the local cohesiveness of personal networks by taking into account the intensity of interactions between local triplets [38]. This measure does not just take into account a number of closed triplets of a node i but also their total relative weight with respect to the total strength of the nodes (see Methods). We also examine how the value of weighted clustering coefficient, averaged over all participants that have attended x events, designated as $\langle c_i^W(x) \rangle$, with the number of attended events. As it is shown in Fig 4, a member's network of personal contacts shows high level of cohesiveness, on the average. Like its non-weighted counterpart, the value of $\langle c_i^W \rangle$ only slightly decreases during member's early involvement in group activities, while later it remains constant and independent of the number of participations. A comparison of the values of weighted and non-weighted clustering coefficients reveals the role of strong relationships in local networks, i.e., whether they form triplets or bridges between different cohesive groups [38]. At the beginning of member's involvement in a group, these two coefficients have similar values, Fig 4, which indicates that the cohesiveness of a subgroup of personal contacts is not that important for the early participation dynamics. As a number of attended events grows, as well as a number and strength of personal contacts, the weighted clustering coefficient becomes larger than its non-weighted counterpart, indicating member's strongest ties with other members who are also friends. The fact that in later

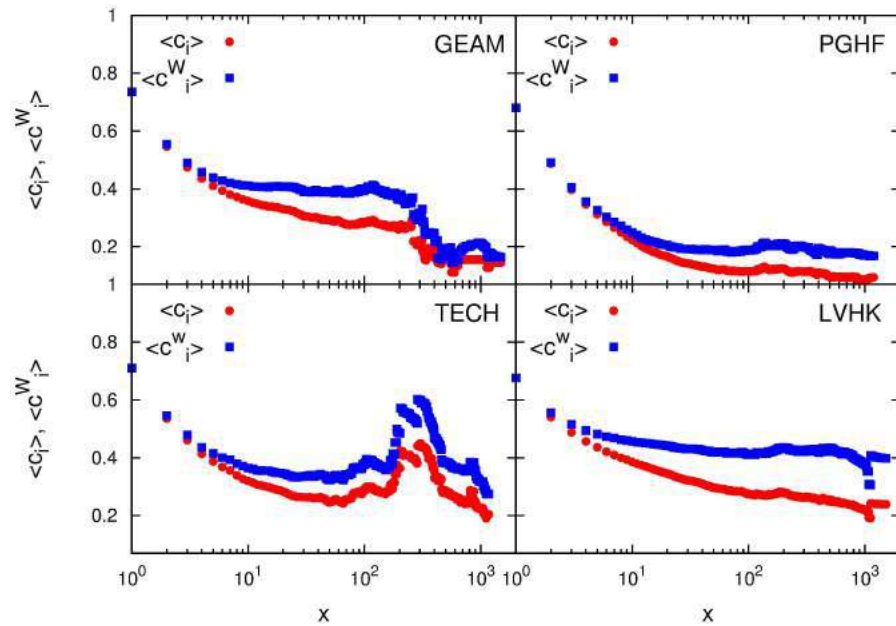


Fig 4. Local cohesiveness of social networks of significant links. Evolution of local cohesiveness of members personal networks, measured by averaged non-weighted $\langle C_i \rangle$ and weighted clustering coefficients $\langle C_i^W \rangle$, with the number of events attended by the member x .

doi:10.1371/journal.pone.0171565.g004

engagement the weighted clustering coefficient is larger than its non-weighted counterpart indicates that the clustering has an important role in the network organization of Meetup groups and thus in the group participation dynamics [38]. Low and very similar values of the clustering and weighted clustering coefficients in networks obtained for randomized data (Fig G in S1 File) further confirm our conclusion about the importance of clustering in the event participation dynamics. The observed discontinuity and decrease of values of the degree, strength and both clustering coefficients, Figs 3 and 4, for groups GEAM and TECH are consequences of a small number of members who attended more than 300 events.

Event importance in group participation dynamics

In our previous work [20], we have shown that the conference participation dynamics is independent of the conference topic, type and size. The same holds true for the Meetup participation dynamics, i.e., the member’s participation patterns in the Meetup group activities do not depend on the group size, category, location or type of activity. However, the size of group events and their time order may influence the structure of network and thus group dynamics. We explore how topological properties of networks, specifically the number of acquaintances and network cohesion, change after the removal of events according to a certain order (see Methods for details).

Firstly, we study how the removal of events according to a certain order influences the number of overall acquaintances in the network. For this purpose we define a measure η (see Methods), which we use to quantify the percentage of the remaining significant acquaintances after the removal of an event. Fig 5 shows the change of measure η after the removal of a fraction r of events according to a chosen strategy. We see that most of new significant connections are usually made during the largest events. The importance of large events for the creation of new acquaintances is especially striking for the groups GEAM, PGHF, and TECH, where

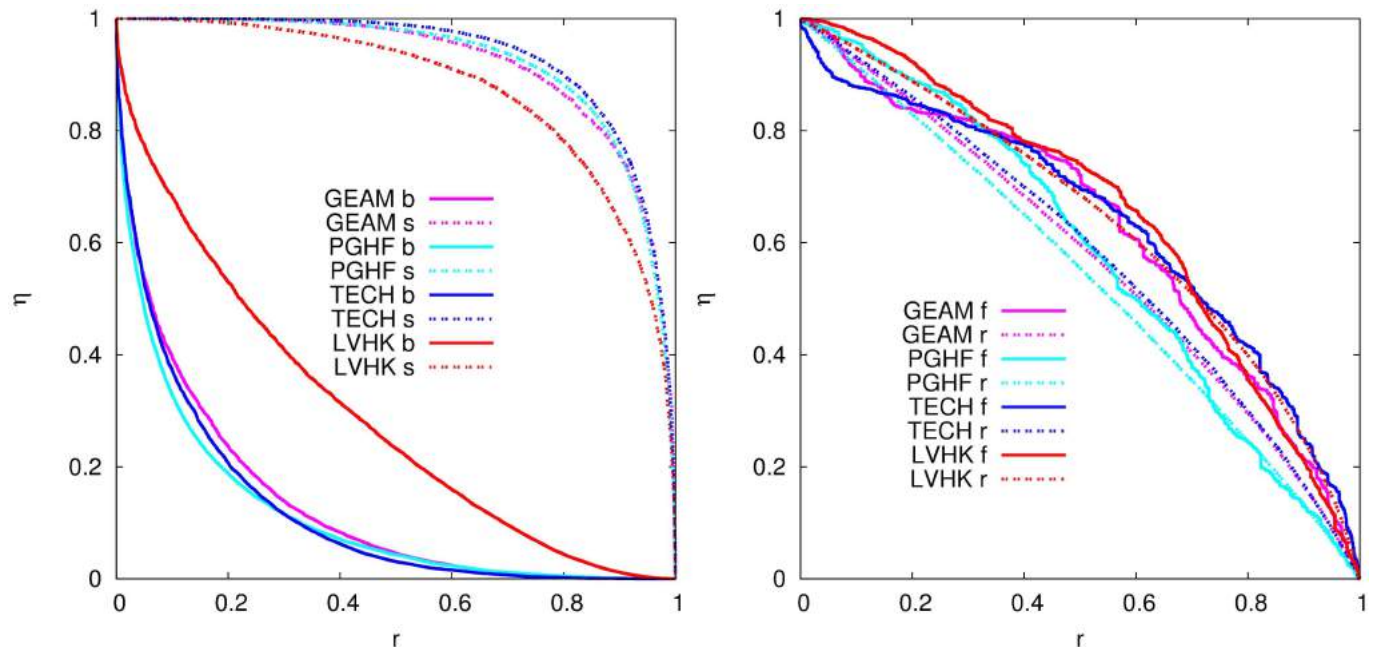


Fig 5. Importance of event size for number of distinctive links in the networks. Change of η with removal of events according to their size (left) and temporal and random order (right). Abbreviations indicate order in which we remove events: **b**—from the largest to the smallest, **s**—from the smallest to the largest, **f**—from the first to the last and **r**—random.

doi:10.1371/journal.pone.0171565.g005

about 80% of acquaintances only met at top 20% of events. For LVHK the decrease is slower, probably due to a difference in the event size fluctuations (see Fig E in S1 File), but still more than 50% of acquaintances disappear if we remove top 40% of events, which is still much higher percentage of contacts compared to random removal of events (see Fig 5 (right)). Similar results are observed when we remove events in the opposite order, Fig 5 (left). Only 20% of acquaintances are being destroyed after the removal of 80% of events, for all four groups. This indicates that new and weak connections are usually formed during large events, while these acquaintances are further strengthened during small meetups. On the other hand, the removal of events according to their temporal order, Fig 5, has very similar effect as random removal, i.e., the value of parameter η decreases gradually as we remove events.

Similar conclusions can be drawn based on the change of average weighted clustering coefficient $\langle C^W \rangle$ (now averaged over all nodes in the network) with the removal of events, Fig 6. Removal of events according to decreasing order of their sizes, does not result in the significant change of $\langle C^W \rangle$. The same value of weighted clustering coefficient, observed even after the removal of 80% of events, shows that small events are not attended by a pair of but rather by a group of old friends. On the other hand, the removal of events in the opposite order results in gradual decrease of $\langle C^W \rangle$. A certain fraction of triads in networks are made by at least one link of low weight. These links are most likely to vanish after the removal of the largest events, which results in the gradual decrease of $\langle C^W \rangle$. Removal of events according to their temporal order results in the change of $\langle C^W \rangle$ similar to one obtained for random removal of events, confirming further that the time ordering of events does not influence the structure of studied networks.

Discussion and conclusion

In this article we explore the event participation dynamics and underlying social mechanism of the Meetup groups. The motivation behind this was to further explore the event driven

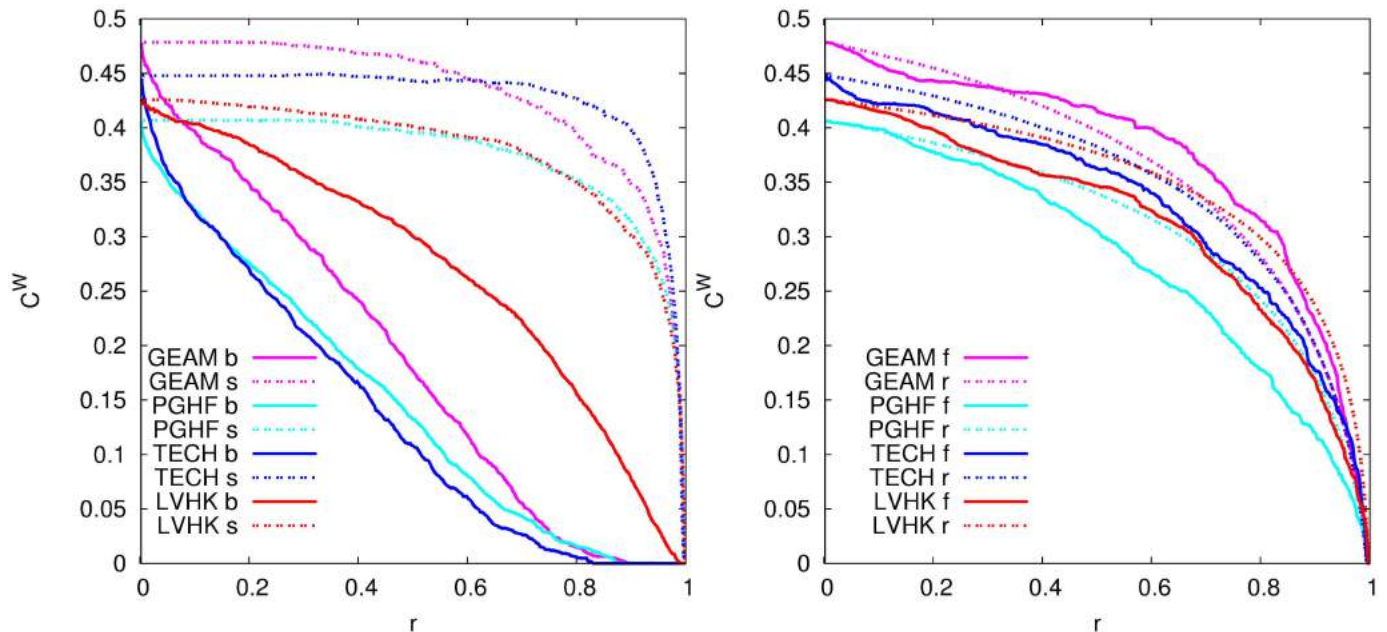


Fig 6. Importance of event size for the network cohesiveness. Change of local network cohesiveness with removal of events according to their size (left) and temporal and random order (right). Abbreviations indicate order in which we remove events: **b**—from the largest to the smallest, **s**—from the smallest to the last and **r**—random.

doi:10.1371/journal.pone.0171565.g006

dynamics, work we have started by exploring participation patterns of scientists at scientific conferences [20], and to better examine the social origins behind the repeated attendance at group events, which was not feasible with the conference data. The results in this manuscript are based on empirical analysis of participation patterns and topological characteristics of networks for four different Meetup groups made up of people who have different motives and readiness to participate in group activities: GEAM, PGHF, TECH, LVHK.

Although these four groups differ in category and type of activity, we have shown that they are all characterized with similar participation patterns: the probability distributions of total number of participations, number of successive participations and time lag between two successive participations follow a power law and truncated power law behavior, with the value of power law exponents between 1 and 3. The resemblance of these patterns to those observed for conference participations [20] indicates that these two, seemingly different, social system dynamics are governed by similar mechanism. This means that the probability for a member to participate in future events depends non-linearly on the balance between the numbers of previous participations and non-participations. As in the case of conferences [20], this behavior is independent of the group category, size, or location, meaning that members association with the community of a Meetup group strongly influence their event participation patterns, and thus the frequency and longevity of their engagement in the group activities.

The Member's association with the community is primarily manifested through her interconnectedness with other members of a specific Meetup group, i.e., in the structure of her personal social network. We have examined topological properties of filtered weighted social networks constructed from the members event co-occurrence. Through network filtering we have emphasized the importance of significant links, the ones which are not the result of coincidence but rather an indicator of social relations. The analysis of local topological properties of these networks has revealed that the strength of connectedness with the community, for the members with small number of participations, is predominantly the consequence of the width

of their social circles. Average strength and degree of members with $q \lesssim 50$, which on average corresponds to only a few participations, are equal, while the strength of members who know more than 50 people and have participated in more than a few events, is several times higher than their degree. This means that after a few participations strengthening of existing ties becomes more important than meeting new people. These arguments are further extended with our observation of the evolution of average strength and degree with the growth of number of participations. Both, average degree and strength, grow, but the growth rate of strength is higher than one of the degree, for all four Meetup groups. All four groups are characterised with very high cohesiveness of their social communities. The evolutions of clustering coefficients, non and weighted one, and their ratio, show that bonding with the community becomes more important as the members' engagement in the group activity progresses. As in the case of conference participations, frequent attendees of group activities tend to form a core whose stability grows with the number of participations [20, 39]. The need of frequent attendees to maintain and increase their bonding with the rest of the community influences their probability to attend future meetings and thus governs the event participation dynamics of the Meetup groups.

The observed structure of personal social networks of the Meetup members is in accordance with previous research on this topic [21–24]. The average size of personal social networks for the most frequent attendees of the Meetup groups GEAM, PGHF, and TECH, is 150 or lower, while the size of the LVHK personal network is less than 500 different connections, i.e., of the same order. This is consistent with the predictions of the Social Brain Hypothesis for the typical human group size. The faster growth of the strength, compared to the one observed for degree, and the constant, non-trivial, value of the clustering coefficients are indicators of the layered structure of social networks. The comparable values of strength and degree, as well as weighted and non-weighted clustering coefficients, observed for small numbers of attendances, indicate that at the beginning all social connections are of the equal importance. As members' engagement with the community grows, she begins to interact with a certain members of the group more often, which results in the non-linear growth of her strength. The higher value of weighted clustering coefficient, compared to its non-weighted counterpart, indicates that member's personal network consists of layers, subgroups of members, characterized with similar strength of mutual relations.

While the group category, type of activity and size do not significantly affect the participation dynamics in the group activities and structure of networks, the size of separate events does have an influence on the evolution of social networks. Large events represent an opportunity for members to make new acquaintances, i.e., to establish new connections. On the other hand, small meetings are typically the gatherings of members with preexisting connections, and their main purpose is to facilitate the stronger bonding among group members. We find that the time order of events is irrelevant for group dynamics.

The universality of the event participation patterns, shown in this and previous work [20], and its socially driven nature give us a better insight not only about the dynamics of studied social communities but also about others which are organised on very similar principles: communities that bring together people with the similar interests and where the participation is voluntary. Having in mind that these type of groups constitute a large part of human life, including all life aspects, understanding their functioning and dynamics is of great importance. Our results not only contribute to the corpus of increasing knowledge, but also indicate the key factor which influences the group longevity and successful functioning: the association of group members with the community. This and recent success stories [40] suggest that complex network theory can be an extremely useful tool in creating successful communities. Future studies will be conducted towards further confirmation of universality of event participation

patterns and better understanding of how social association and contacts can be used for creating conditions for successful functioning of learning and health support groups.

Materials and methods

Data

There are more than 240000 groups in 181 countries classified into 33 categories active in the Meetup community [41]. For each of selected four groups, we have used the Meetup public API to access the data and collect the list of events organized by the group and the information on the members who confirmed their participation (RSVP) in the given event since the group's beginnings. Each member has a unique id which enables us to follow her activity in the group events during the time. The collected data have been fully anonymised and we did not collect any personal information about the group members. We have complied with terms of use of Meetup website. More details about the group sizes and the number of events is given in [Table 1](#).

Network construction and filtering

Network construction. We start with a bipartite member event network, which we represent with participation matrix B . Let N_m denotes total number of members in the group and N_e is total number of events organized by the group. If the member i participated in the event l element of matrix B_{il} takes a value 1, otherwise $B_{il} = 0$. In the bipartite network created in this way, members' degree is equal to total number of events member participated in, while events' degree is defined as total number of members that have attended that event. The social network, which is the result of members interactions during the Meetup events and is represented by weighted matrix W , is created from the weighted projection of bipartite network to members partition [42, 43]. In the obtained weighted network nodes correspond to individual members while the value of the element of weighted matrix W_{ij} corresponds to number of common events two members have attended together.

Network filtering. The observed weighted network is dense network where some of the non-zero edges can be the result of coincidence. For instance, these edges can be found between members who attended large number of events or events with many participants, and therefore they do not necessarily indicate social connections between members. The pruning of these type of networks and separation of significant edges from non-significant ones is not a trivial task [36, 37, 44]. For this reasons we start from bipartite network and use method that determines the significance of W_{ij} link based on configuration model of random bipartite networks [36, 37, 45, 46]. In this model of random networks the event size and the number of events a member attended are fixed, while all other correlations are destroyed (see SI for further explanations). Based on this model, for each link in bipartite network, B_{il} , we determine the probability p_{il} that user i has attended event l . The assumption of uncorrelated network enables us to also estimate the probability that two members, i and j , have attended the same event, which is equal to $p_{il}p_{jl}$. Probability that two members have attended the same w events is then given by Poisson binomial distribution

$$P_{ij}(w) = \sum_{M_w} \prod_{l \in M_w} p_{il}p_{jl} \prod_{l \notin M_w} (1 - p_{il}p_{jl}) \tag{1}$$

where M_w is the subset of w events that can be chosen from given M events [36, 37, 47]. We define p -value as probability that two members i and j has co-occurred on at least w_{ij} events,

i.e., that the link weight between these two members is w_{ij} or higher

$$p\text{-value}(w_{ij}) = \sum_{w \geq w_{ij}} P_{ij}(w). \tag{2}$$

The relationship between users i and j will be considered statistically significant if $p\text{-value}(w_{ij}) \leq p_{trs}$. In our case, threshold $p_{trs} = 0.05$. All links with $p\text{-value}(w_{ij}) > p_{trs}$ are consequence of chance and are considered as non-significant and thus removed from the network. This way we obtain weighted social network of significant relations between members of the Meetup group W_{ij}^S . The details on how we estimate p_{ij} and $P_{ij}(w)$ for each link are given in SI.

Topological measures. All topological measures considered in this work are calculated for weighted social network of significant relations W_{ij}^S . We consider the following topological measures of the nodes:

- The node degree $q_i = \sum_j \mathcal{H}(W_{ij}^S)$, where \mathcal{H} is Heaviside function ($\mathcal{H}(x) = 1$ if $x > 0$ otherwise $\mathcal{H}(x) = 0$);
- The node strength $s_i = \sum_j W_{ij}^S$ [7];
- Non-weighted clustering coefficient of the node $c_i = \frac{1}{q_i(q_i-1)} \sum_{j,m} \mathcal{H}(W_{ij}^S) \mathcal{H}(W_{im}^S) \mathcal{H}(W_{jm}^S)$ [7].
- Weighted clustering coefficient of the node $c_i^W = \frac{1}{s_i(q_i-1)} \sum_{j,m} \frac{W_{ij}^S + W_{im}^S}{2} \mathcal{H}(W_{ij}^S) \mathcal{H}(W_{im}^S) \mathcal{H}(W_{jm}^S)$ [38].

Weighted clustering coefficient of the network $\langle C^W \rangle$ and its non-weighted counterpart $\langle C \rangle$ are values averaged over all nodes in the network.

The event relevance

In order to explore the relevance of event size and time ordering for the evolution of social network topology we analyze how removal of events, according to specific ordering, influences the number of acquaintance and network cohesion. Specifically, we observe change of measure η , which represents the fraction of the remaining acquaintances, and weighted clustering coefficient $\langle C^W \rangle$ after the removal of a fraction r of events. The removal of event results in change of link weights between group members. For instance, if two members, i and j , have participated in event l , the removal of this event will result in the decrease of the link weight W_{ij}^S by one. Further removal of events in which these two members have co-occurred will eventually lead to termination of their social connection, i.e., $W_{ij}^S = 0$. If $W^S(r)$ is the matrix of link weights after the removal of a fraction r of events and W^S is the original matrix of significant relations, then the value of parameter η after the removal of r events is calculated as

$$\eta(r) = \frac{\sum_{ij} \mathcal{H}(W_{ij}^S(r))}{\sum_{ij} \mathcal{H}(W_{ij}^S)}, \tag{3}$$

The value of weighted clustering coefficient $\langle C^W \rangle$ after the removal of a fraction r of events is calculated using the same formula as for the $\langle C^W \rangle$ just using the value of $W^S(r)$ instead of W^S .

We remove events according to several different strategies:

- We sort events according to their size. Then, we remove sorted events in descending and ascending order.
- We remove events according to their time-order, from the first to the last.

- We remove events in random order. We perform this procedure for each list of events 100 times.

Supporting information

S1 File. Supplementary information: Associative nature of event participation dynamics: a network theory approach. The probability distribution $P(x)$ of total numbers of participations in group events x , obtained from the empirical data for the four selected Meetup groups (blue circles). We also show truncated power law fit $x^{-\alpha}e^{-Bx}$ (solid lines), power law fit $x^{-\gamma}$ (dotted-dashed lines), and exponential fit $e^{-\lambda x}$ (dotted lines). **Fig A** Log likelihood ratio \mathcal{R} and the π -value compare fits to the power law and fits to the truncated power law for the probability distribution of total numbers of participations in group events. **Table A** The probability distribution of successive numbers of participations in group events x_s , for the four selected Meetup groups. The probability distribution follows power law behavior $P(x_s) \sim x_s^{-\gamma}$. **Fig B** The probability distribution of time lags between two successive participations in group events y_s , for the four selected Meetup groups. The probability distribution follows truncated power law behavior $P(y_s) \sim y_s^{-\alpha}e^{-By_s}$. **Fig C** The probability distribution of link weights in a weighted network before and after filtering, for the four selected Meetup groups. **Fig D** The dependence of a degree strength ratio on the number of participations, averaged over all members for the four considered Meetup groups. Red circles correspond to results obtained from empirical data, while blue squares correspond to randomized data. **Fig E** The dependence of group members' average degree $\langle q \rangle$ and strength $\langle s \rangle$ on numbers of participations for a real weighted network and a randomized network. **Fig F** The dependence of group members' average non-weighted $\langle c_i \rangle$ and weighted clustering coefficient $\langle c_i^W \rangle$ on numbers of participations for a real weighted network and a randomized network. **Fig G** The probability distribution of relative size fluctuations $\frac{\langle e \rangle - e}{\langle e \rangle}$, for the four considered Meetup groups, where e is the event size and $\langle e \rangle$ is the average event size. **Fig H.**
(PDF)

Acknowledgments

Numerical simulations were run on the PARADOX supercomputing facility at the Scientific Computing Laboratory of the Institute of Physics Belgrade.

Author contributions

Conceptualization: JS MMD.

Data curation: JS.

Formal analysis: JS MMD.

Investigation: JS.

Methodology: JS MMD.

Visualization: JS MMD.

Writing – original draft: JS MMD.

References

1. Castellano C, Fortunato S, Loreto V. Statistical physics of social dynamics. *Rev Mod Phys.* 2009; 81:591–646. doi: [10.1103/RevModPhys.81.591](https://doi.org/10.1103/RevModPhys.81.591)

2. Nowak MA. Five Rules for the Evolution of Cooperation. *Science*. 2006; 314(5805):1560–1563. doi: [10.1126/science.1133755](https://doi.org/10.1126/science.1133755) PMID: [17158317](https://pubmed.ncbi.nlm.nih.gov/17158317/)
3. Fowler JH, Christakis NA. Cooperative behavior cascades in human social networks. *Proc Natl Acad Sci USA*. 2010; 107(12):5334–5338. doi: [10.1073/pnas.0913149107](https://doi.org/10.1073/pnas.0913149107) PMID: [20212120](https://pubmed.ncbi.nlm.nih.gov/20212120/)
4. Granovetter MS. The Strength of Weak Ties. *Am J Sociol*. 1973; 78(6):1360–1380. doi: [10.1086/225469](https://doi.org/10.1086/225469)
5. Pastor-Satorras R, Castellano C, Van Mieghem P, Vespignani A. Epidemic processes in complex networks. *Rev Mod Phys*. 2015; 87:925–979. doi: [10.1103/RevModPhys.87.925](https://doi.org/10.1103/RevModPhys.87.925)
6. Mitrović Dankulov M, Melnik R, Tadić B. The dynamics of meaningful social interactions and the emergence of collective knowledge. *Sci Rep*. 2015; 5:12197. doi: [10.1038/srep12197](https://doi.org/10.1038/srep12197)
7. Boccaletti S, Latora V, Moreno Y, Chavez M, Hwang DU. Complex networks: Structure and dynamics. *Phys Rep*. 2006; 424(4–5):175–308. doi: [10.1016/j.physrep.2005.10.009](https://doi.org/10.1016/j.physrep.2005.10.009)
8. Holme P, Saramäki J. Temporal networks. *Phys Rep*. 2012; 519(3):97–125. doi: [10.1016/j.physrep.2012.03.001](https://doi.org/10.1016/j.physrep.2012.03.001)
9. Aral S, Walker D. Identifying Influential and Susceptible Members of Social Networks. *Science*. 2012; 337(6092):337–341. doi: [10.1126/science.1215842](https://doi.org/10.1126/science.1215842) PMID: [22722253](https://pubmed.ncbi.nlm.nih.gov/22722253/)
10. González-Bailón S, Borge-Holthoefer J, Moreno Y. Broadcasters and Hidden Influentials in Online Protest Diffusion. *Am Behav Sci*. 2013; 57(7):943–965. doi: [10.1177/0002764213479371](https://doi.org/10.1177/0002764213479371)
11. Lin YR, Chi Y, Zhu S, Sundaram H, Tseng BL. Facetnet: A Framework for Analyzing Communities and Their Evolutions in Dynamic Networks. In: *Proceedings of the 17th International Conference on World Wide Web. WWW'08; 2008*. p. 685–694.
12. Mitrović M, Paltoglou G, Tadić B. Quantitative analysis of bloggers' collective behavior powered by emotions. *J Stat Mech*. 2011; 2011(02):P02005.
13. Garas A, Garcia D, Skowron M, Schweitzer F. Emotional persistence in online chatting communities. *Sci Rep*. 2012; 2:402. doi: [10.1038/srep00402](https://doi.org/10.1038/srep00402) PMID: [22577512](https://pubmed.ncbi.nlm.nih.gov/22577512/)
14. Török J, Iñiguez G, Yasserli T, San Miguel M, Kaski K, Kertész J. Opinions, Conflicts, and Consensus: Modeling Social Dynamics in a Collaborative Environment. *Phys Rev Lett*. 2013; 110:088701. doi: [10.1103/PhysRevLett.110.088701](https://doi.org/10.1103/PhysRevLett.110.088701) PMID: [23473207](https://pubmed.ncbi.nlm.nih.gov/23473207/)
15. Yasserli T, Sumi R, Rung A, Kornai A, Kertész J. Dynamics of Conflicts in Wikipedia. *PLoS ONE*. 2012; 7(6):1–12. doi: [10.1371/journal.pone.0038869](https://doi.org/10.1371/journal.pone.0038869)
16. Montazeri A, Jarvandi S, Soghraand Haghighat, Vahdani A, Mariamand Sajadian, Ebrahimi M, Haji-Mahmoodi M. Anxiety and depression in breast cancer patients before and after participation in a cancer support group. *Patient Educ Couns*. 2001; 45:195–198. doi: [10.1016/S0738-3991\(01\)00121-5](https://doi.org/10.1016/S0738-3991(01)00121-5) PMID: [11722855](https://pubmed.ncbi.nlm.nih.gov/11722855/)
17. Davison KP, Pennebaker JW, Dickerson SS. Who talks? The social psychology of illness support groups. *Am Psychol*. 2000; 55:205–217. doi: [10.1037/0003-066X.55.2.205](https://doi.org/10.1037/0003-066X.55.2.205) PMID: [10717968](https://pubmed.ncbi.nlm.nih.gov/10717968/)
18. Tam Cho WK, Gimpel JG, Shaw DR. The Tea Party Movement and the Geography of Collective Action. *Q J Polit Sci*. 2012; 7:105–133. doi: [10.1561/100.00011051](https://doi.org/10.1561/100.00011051)
19. Weinberg BD, Williams CB. The 2004 US Presidential campaign: Impact of hybrid offline and online 'meetup' communities. *J Direct Data Digit Mark Pract*. 2006; 8(1):46–57. doi: [10.1057/palgrave.ddmp.4340552](https://doi.org/10.1057/palgrave.ddmp.4340552)
20. Smiljanić J, Chatterjee A, Kauppinen T, Mitrović Dankulov M. A Theoretical Model for the Associative Nature of Conference Participation. *PLoS ONE*. 2016; 11(2):1–12.
21. Dunbar RIM. Coevolution of neocortical size, group size and language in humans. *Behavioral and Brain Sciences*. 1993; 16(4):681–694. doi: [10.1017/S0140525X00032325](https://doi.org/10.1017/S0140525X00032325)
22. Dunbar RIM. Mind the Gap: Or Why Humans Aren't Just Great Apes. *Proceedings of the British Academy*. 2008; 154:403–423.
23. Hill RA, Dunbar RIM. Social network size in humans. *Human Nature*. 2003; 14(1):53–72. doi: [10.1007/s12110-003-1016-y](https://doi.org/10.1007/s12110-003-1016-y) PMID: [26189988](https://pubmed.ncbi.nlm.nih.gov/26189988/)
24. Dunbar RIM. Constraints on the evolution of social institutions and their implications for information flow. *Journal of Institutional Economics*. 2011; 7(3):345–371. doi: [10.1017/S1744137410000366](https://doi.org/10.1017/S1744137410000366)
25. Sessions LF. How offline gatherings affect online communities. *Information, Communication & Society*. 2010; 13(3):375–395. doi: [10.1080/13691180903468954](https://doi.org/10.1080/13691180903468954)
26. Hristova D, Quattrone G, Mashhadi A, Capra L. The Life of the Party: Impact of Social Mapping in OpenStreetMap. In: *Proceedings of the Seventh International AAAI Conference on Web and Social Media. ICWSM'13; 2013*. p. 234–243.

27. Qiao Z, Zhang P, Zhou C, Cao Y, Guo L, Zhang Y. Event Recommendation in Event-based Social Networks. In: Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence. AAAI'14; 2014. p. 3130–3131.
28. Zhang W, Wang J, Feng W. Combining Latent Factor Model with Location Features for Event-based Group Recommendation. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'13; 2013. p. 910–918.
29. Pham TAN, Li X, Cong G, Zhang Z. A general graph-based model for recommendation in event-based social networks. In: 2015 IEEE 31st International Conference on Data Engineering; 2015. p. 567–578.
30. Macedo AQ, Marinho LB, Santos RLT. Context-Aware Event Recommendation in Event-based Social Networks. In: Proceedings of the 9th ACM Conference on Recommender Systems. RecSys'15; 2015. p. 123–130.
31. Liu X, He Q, Tian Y, Lee WC, McPherson J, Han J. Event-based Social Networks: Linking the Online and Offline Social Worlds. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'12; 2012. p. 1032–1040.
32. Jiang JY, Li CT. Analyzing Social Event Participants for a Single Organizer. In: International AAAI Conference on Web and Social Media; 2016. p. 599–602.
33. McCully W, Lampe C, Sarkar C, Velasquez A, Sreevivasan A. Online and Offline Interactions in Online Communities. In: Proceedings of the 7th International Symposium on Wikis and Open Collaboration. WikiSym'11; 2011. p. 39–48.
34. Palla G, Barabási AL, Vicsek T. Quantifying social group evolution. *Nature*. 2007; 446(7136):664–667. doi: [10.1038/nature05670](https://doi.org/10.1038/nature05670) PMID: [17410175](https://pubmed.ncbi.nlm.nih.gov/17410175/)
35. Backstrom L, Huttenlocher D, Kleinberg J, Lan X. Group Formation in Large Social Networks: Membership, Growth, and Evolution. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD'06; 2006. p. 44–54.
36. Dianati N. A maximum entropy approach to separating noise from signal in bimodal affiliation networks. *ArXiv e-prints*. 2016;.
37. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Grandcanonical projection of bipartite networks. *ArXiv e-prints*. 2016;.
38. Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A. The architecture of complex weighted networks. *Proc Natl Acad Sci USA*. 2004; 101(11):3747–3752. doi: [10.1073/pnas.0400087101](https://doi.org/10.1073/pnas.0400087101) PMID: [15007165](https://pubmed.ncbi.nlm.nih.gov/15007165/)
39. Van Dijk J, Maier G. ERSa Conference participation: does location matter? *Pap Reg Sci*. 2006; 85(4):483–504. doi: [10.1111/j.1435-5957.2006.00102.x](https://doi.org/10.1111/j.1435-5957.2006.00102.x)
40. Cosgrave P. Engineering Serendipity: The Story of Web Summit's Growth; 2014. Available from: <https://goo.gl/H3aWMI>.
41. Meetup Datasets;. Available from: <https://www.meetup.com/>.
42. Mitrović M, Tadić B. Bloggers behavior and emergent communities in Blog space. *Eur Phys J B*. 2010; 73(2):293–301. doi: [10.1140/epjb/e2009-00431-9](https://doi.org/10.1140/epjb/e2009-00431-9)
43. Mitrović M, Paltoglou G, Tadić B. Networks and emotion-driven user communities at popular blogs. *Eur Phys J B*. 2010; 77(4):597–609. doi: [10.1140/epjb/e2010-00279-x](https://doi.org/10.1140/epjb/e2010-00279-x)
44. Dianati N. Unwinding the hairball graph: Pruning algorithms for weighted complex networks. *Phys Rev E*. 2016; 93:012304. doi: [10.1103/PhysRevE.93.012304](https://doi.org/10.1103/PhysRevE.93.012304) PMID: [26871089](https://pubmed.ncbi.nlm.nih.gov/26871089/)
45. Saracco F, Di Clemente R, Gabrielli A, Squartini T. Randomizing bipartite networks: the case of the World Trade Web. *Sci Rep*. 2015; 5:10595. doi: [10.1038/srep10595](https://doi.org/10.1038/srep10595) PMID: [26029820](https://pubmed.ncbi.nlm.nih.gov/26029820/)
46. Cellai D, Bianconi G. Multiplex networks with heterogeneous activities of the nodes. *Phys Rev E*. 2016; 93:032302. doi: [10.1103/PhysRevE.93.032302](https://doi.org/10.1103/PhysRevE.93.032302) PMID: [27078361](https://pubmed.ncbi.nlm.nih.gov/27078361/)
47. Liebig J, Rao A. Fast extraction of the backbone of projected bipartite networks to aid community detection. *Europhys Lett*. 2016; 113(2):28003. doi: [10.1209/0295-5075/113/28003](https://doi.org/10.1209/0295-5075/113/28003)