AFRL-OSR-VA-TR-2015-0212

DARPA EnsembleBased Modeling Large Graphs & Applications to Social Networks

Zoltan Toroczkai
UNIVERSITY OF NOTRE DAME DU LAC

07/29/2015
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
29-07-2015	Final Report	Aug 01 2012 - July 31, 2015

4. TITLE AND SUBTITLE
DARPA Ensemble-Based Modeling Large Graphs & Applications to Social Networks

5a. CONTRACT NUMBER
FA9550-12-1-0405

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Dr. Zoltan Toroczkai,
Department of Physics, 225 NSH,
University of Notre Dame,
Notre Dame, IN, 46556

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Notre Dame Du Lac,
Office of Research
940 Grace Hall, Notre Dame, IN 46556-5602

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research, 875 N. Randolph St. Room 3112, Arlington, VA, 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR, DARPA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A - Approved for Public release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Economies, social and political systems all exist embedded within complex and dynamic networks. Understanding the network landscape is a critical necessity for winning conflicts and securing global safety, peace and stability. It helps prevent surprises and provides both a tactical and a strategic upper hand in our interactions both with our allies and our adversaries. Understanding complex networks requires the development of reliable mathematical and computational tools that efficiently probe the data and extract the relevant information. This is a very difficult undertaking and it requires marshaling methods from traditionally disparate areas, including graph theory (discrete mathematics), statistical physics, statistics, theoretical computer science, algorithm development, and data mining in a sustained fashion. This project has been developing mathematical and computational methods pushing the state of the art in data driven modeling of complex networks. The algorithms and the methods developed here have been validated against real-world network datasets.

15. SUBJECT TERMS
Complex networks, mathematical modeling, data driven analysis and prediction

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER *(Include area code)*
UU	UU	UU	SAR	58	

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18
Adobe Professional 7.0

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATES COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33615-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report, e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5d. PROJECT NUMBER. Enter all project numbers as they appear in the report, e.g. 1F665702D1257; ILIR.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Report: Ensemble-Based Modeling of Large Graphs and its Applications to Social Networks

Focus Area One: The structure and dynamics of large graphs

Award No: FA9550-12-1-0405
PI: Zoltan Toroczkai, University of Notre-Dame

As detailed in our proposal and the SOWs, the work on this project was performed following three main directions, or task areas: **T1.** *Constrained graph ensembles*, **T2.** *Finding structures of interest to influence and control networks*, **T3.** *Data-driven m*ethods.

This is a condensed summary of the work done under this project. The full description of the results are found in the 47 papers that have been published or submitted for publication as enlisted in the output section of this report.

Overall goal. The project's main goal was to develop systematic and principled mathematical approaches to modeling complex networks of DARPA's interest. The methods and algorithms (exact or heuristic) have been validated with direct applications on datasets, which include: brain network data, social interactions data (Enron, Facebook, Wikipedia, Mobile phones, etc), and large-scale infrastructure data (US roadways).

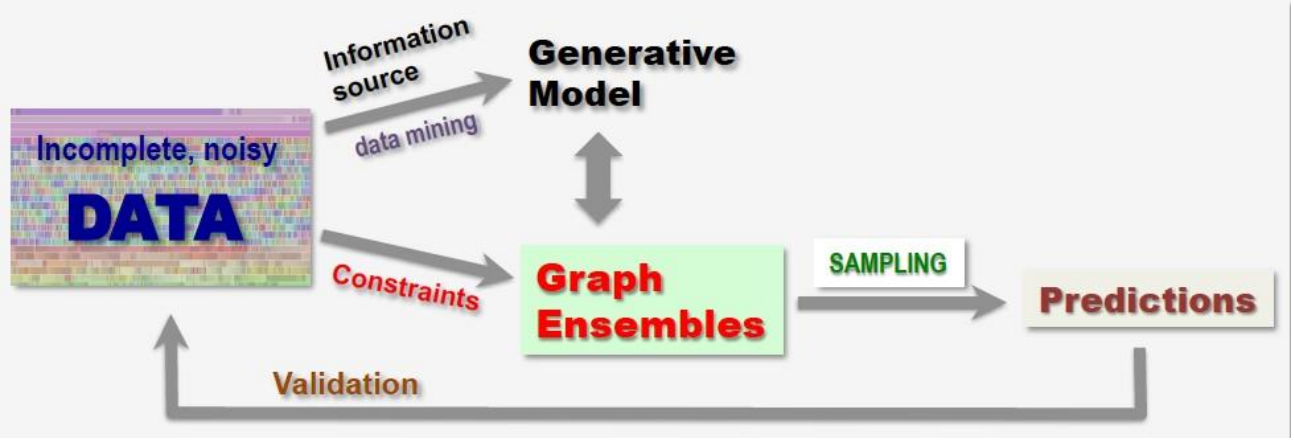


Figure 1. A pictorial overview of the project's components and their interactions.

This project has achieved all of its major goals that were set up in the proposal. However, as with any fundamental research, our results have raised new exciting and important questions, both fundamental and application focused, whose answer would be valuable to DARPA and network science in general. We are looking forward to the possibility of exploring these as part of new funding initiatives.

High Level Summary

Modeling and predicting the structure and behavior of real-world complex networks, including those with defense applications requires a suite of computational and mathematical tools that need to address model generation, network search and data analysis problems. During this project we have achieved our main proposed goals in all these research fronts. **T1.** *Modeling with Constrained Graph*

Ensembles. We provided results to characterize graph ensembles defined by empirical data in form of sharp constraints, including existence, construction, sampling and graph counting problems. Developed novel proofs for the Markov Chain Monte Carlo mixing time problem for graph sampling based on degrees and joint degrees. For Exponential Random Graph Models we provided an understanding of the degeneracy problem and proposed a novel method that eliminates this problem. Described spectral properties of random geometric graph ensembles in hyperbolic spaces and connected them to real social networks. **T2.** *Finding Structures of Interest*. Shown that minimum dominating sets (MDS) provide an effective way to search, monitor and influence large networks. Developed linearly scalable heuristic algorithms based on the probabilistic method (Lovasz Local Lemma) to identify MDS in large networks; we demonstrated its applicability on a model for opinion spread. Expanded the setting for the applicability of LLL for network problems to be also used as a counting tool. **T3.** *Data Driven Methods*. Developed novel results for Mixed Orthogonal Arrays to be used for the design of experiments and statistical queries of large networks. Developed novel tools for link prediction, node prominence prediction, and influence propagation in multirelational networks. Modeled co-evolutionary processes for subgraph – embedding graph relationships.

DoD Relevance

Economies, social and political systems all exist embedded within complex and dynamic networks. Understanding the network landscape is a critical necessity for winning conflicts and securing global safety, peace and stability. It helps prevent surprises and provides both a tactical and a strategic upper hand in our interactions both with our allies and our adversaries. Understanding complex networks requires the development of reliable mathematical and computational tools that efficiently probe the data and extract the relevant information. This is a very difficult undertaking and it requires marshaling methods from traditionally disparate areas, including graph theory (discrete mathematics), statistical physics, statistics, theoretical computer science, algorithm development, and data mining in a sustained fashion. This project has been developing mathematical and computational methods pushing the state of the art in data driven modeling of complex networks. The algorithms and the methods developed here have been validated against real-world network datasets.

Output

The output from this research is in form of publications, preprints, presentations at conferences and meetings. A total of **47 publications** and preprints with another 13 publications under preparation, **7 PhDs** and **1 MSc theses** generated. Other highlights include students graduated (8), faculty awards (2), conference presentations (69) and other activities and interactions (such as conferences organized), see the sections below.

Publications

[p1] C. Orsini, M.M. Dankulov, A. Jamakovic, P. Madahevan, P. Colomer-de-Simon, A. Vahdat, **K.E. Bassler**, **Z. Toroczkai**, M. Boguna, G. Caldarelli, S. Fortunato, and D. Krioukov. How random are complex networks. *Nature Communications*, submitted (2015). http://arxiv.org/abs/1505.07503

[p2] **I. Miklos**, **H. Smith**. Sampling and counting genome rearrangement scenarios. Accepted and to

appear in 13th *RECOMB Satellite Conference on Comparative Genomics (RECOMB-CG 2015)* to be held in Frankfurt, Germany, from October 4 to 7, 2015.

[p3] **Z. Toroczkai**, M. Ercsey-Ravasz, R. Gămănuț, **Sz. Horvát**, L. Magrou, B. Gămănuț, A. Burkhalter, D. C. Van Essen, K. Knoblauch, H. Kennedy. Spatial Embedding and Wiring Cost Constrain the Functional layout of the Cortical Network. *Nature*, submitted (2015).

[p4] **K.E. Bassler**, D. Dhar, and R.K.P. Zia. Networks with preferred degree: A mini-review and some new results. *J. Stat. Mech.* P07013 (2015). http://arxiv.org/abs/1506.05688

[p5] T. Paixao, **K.E. Bassler**, and R. Azevedo. Emergent speciation by multiple Dobzhansky-Muller incompatibilities. PLoS Comp. Biol *submitted* (2015).

[p6] **I. Miklos**, **H. Smith**, The computational complexity of calculating partition functions of optimal medians with Hamming distance. Submitted, http://arxiv.org/abs/1506.06107

[p7] R. Chauhan, P. Datta, S. Trevino III, D. Schnappinger, **K.E. Bassler**, G. Balazsi, and M.L. Gennaro. Reconstruction and topological features of the sigma factor network of Mycobacterium tuberculosis. *Nature Comm.* submitted (2015).

[p8] **Sz. Horvát**, **É. Czabarka** & **Z. Toroczkai**. Reducing Degeneracy in Maximum Entropy Models of Networks. *Physical Review Letters*, **114**, 158701 (2015). http://arxiv.org/abs/1407.0991

[p9] **K.E. Bassler**, C.I. Del Genio, **P.L. Erdos**, **I. Miklos**, **Z. Toroczkai**. Exact sampling of graphs with prescribed degree correlations. *New Journal of Physics*, in press (2015). http://arxiv.org/abs/1503.06725

[p10] **J. Xu**, T.L. Wickramarathne, **N.V. Chawla**, Representing Higher Order Dependencies in Networks, *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, submitted (2015).

[p11] **H. Smith**, **L.A. Szekely**, Hua Wang, Eccentricity in trees, *Electr. J. Comb.* submitted (2015). http://arxiv.org/abs/1408.5865

[p12] **P.L. Erdős**, S.Z. Kiss, **I. Miklós**, & L. Soukup: Approximate Counting of Graphical Realizations, *PLoS ONE*, **10**(7), e0131300 (2015).

[p13] **Y. Yang**, **N.V. Chawla**, R.N. Lichtenwalter, **Y. Dong**, Influence Activation Model: A New Perspective in Social Influence Analysis and Social Network Evolution, *Science Advances*, submitted (2015).

[p14] L. Lu and **L. A. Székely**. A new asymptotic enumeration technique: the Lovasz Local Lemma. *J. Comb. Theor. Ser. B*, final version accepted (2015). http://arxiv.org/abs/0905.3983

[p15] **E. Czabarka**, **A. Dutle**, T. Johnston, **L. A. Szekely**, Abelian groups yield many large families for the diamond problem, *Europ. J. Math.* submitted (2015).

[p16] **Y. Dong**, J. Zhang, J. Tang, **N. V. Chawla**, B. Wang. Coupled Link Prediction in Networks. *21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, submitted, (2015).

[p17] **Y.Dong**, F. Pinelli, F. Calabrese, **N.V. Chawla**. Inferring Unusual Crowd Events From Mobile Phone Call Detail Records. *European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Database*, submitted (2015).

[p18] **P.L. Erdős**, **I. Miklós** & **Z. Toroczkai**. A decomposition based proof for fast mixing of a Markov Chain over balanced realizations of a Joint Degree Matrix. *SIAM Discr. Math.* **29**, 481-499, (2015). http://arxiv.org/abs/1307.5295

[p19] **K.E. Bassler**, W. Liu, B. Schmittmann, and R.K.P. Zia, Extreme Thouless effect in a minimal model of dynamic social networks, *Phys. Rev. E*, **91**, 042012 (2015).

[p20] Y. Dong, J. Tang, **N.V. Chawla**, T. Lou, **Y. Yang**, B. Wang, Inferring social status and rich club effects in enterprise communication networks, *PLOS One*, **10**(3), e0119446 (2015). http://arxiv.org/abs/1404.3708v3

[p21] **A. Nyberg**, T. Gross, and **K.E. Bassler**, Mesoscopic structures and the Laplacian spectra of Random Geometric Graphs, *J. Complex Networks* cnv004 (2015). doi: 10.1093/comnet/cnv004

[p22] K. Knoblauch, M. Ercsey-Ravasz, H. Kennedy and **Z. Toroczkai**. The Brain in Space. Chapter in *The 22nd Colloque Médecine et Recherche of the Fondation Ipsen in the Neurosciences series: "Micro-, meso- and macro-connectomics of the brain" Fondation IPSEN*, Paris, France. Eds: H. Kennedy, D. Van Essen, Y. Christen. in press Springer, Heidelberg (2015).

[p23] **S. Trevino III**, **A. Nyberg**, C.I. Del Genio, and **K.E. Bassler**, Fast and accurate determination of modularity and its effect size. *J. Stat. Mech.* P02003, (2015).

[p24] **E. Czabarka**, **A. Dutle**, **P.L. Erdos**, & **I. Miklos**. On realizations of a Joint Degree Matrix. *Disc. Appl. Math.* **181**, 283-288 (2015) http://arxiv.org/abs/1302.3548

[p25] Y. Dong, R. A. Johnson, **N. V. Chawla**, Will This Paper Increase Your *h*-index? Scientific Impact Prediction, In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining WSDM'15*, pp. 149-158. (2015) http://dx.doi.org/10.1145/2684822.2685314. (Best Paper Award Nomination 4/39)

[p26] **F. Molnár Jr.**, **N. Derzsy**, B. K. Szymanski, and **G. Korniss**, Building Damage-Resilient Dominating Sets in Complex Networks against Random and Targeted Attacks. *Scientific Reports* **5**, 8321 (2015). http://dx.doi.org/10.1038/srep08321

[p27] Y. Ren, M. Ercsey-Ravasz, P. Wang, M. C. Gonzalez & **Z. Toroczkai**. Predicting commuter flows in spatial networks using a radiation model based on temporal ranges. *Nature Communications*, **5**, 5347 (2014) | http://arxiv.org/abs/1410.4849

[p28] **L.A. Székely**, H. Wang. Extremal values of ratios: distance problems vs. subtree problems in trees II. *Discr. Math.* **322**, 36-47 (2014).

[p29] Y. Dong, Y. Yang, J. Tang, Y. Yang, and N. V. Chawla. 2014. Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '14). ACM, New York, NY, USA, pp. 15-24. http://doi.acm.org/10.1145/2623330.2623703

[p30] **F. Molnar Jr.**, **N. Derzsy**, **E. Czabarka**, **L.A. Székely,** B.K. Szymanski & **G. Korniss**. Dominating scale-free networks using generalized probabilistic methods. *Scientific Reports* **4**, 6308 (2014). http://dx.doi.org/10.1038/srep06308

[p31] Y. Yang, Y. Dong, and **N.V. Chawla**, Predicting Node Degree Centrality with the Node Prominence Profile. *Scientific Reports* **4**, 7236 (2014); http://dx.doi.org/10.1038/srep07236

[p32] S. Hossein, M.D. Reichl, and **K.E. Bassler**. Symmetry in Critical Random Boolean Network Dynamics. *Phys. Rev. E* **89**, 042808 (2014).

[p33] X. Jian, T.L. Wickramarathne, N.V. Chawla, E.K. Grey, K. Steinhaeuser, R.P. Keller, J.M. Drake, and D.M. Lodge. Improving management of aquatic invasions by integrating shipping network, ecological, and environmental data: Data mining for social good. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1699-1708. ACM, 2014. http://dl.acm.org/citation.cfm?id=2623364

[p34] R.N. Lichtenwalter, **N.V. Chawla**. Vertex collocation profiles: theory, computation, and results. *SpringerPlus* **3**(1), 116 (2014) | http://dx.doi.org/10.1186/2193-1801-3-116

[p35] **L.A. Székely**, H. Wang. Extremal values of ratios: distance problems vs. subtree problems in trees *Electr. J. Comb.* **20**(1), #P67, (2013).

[p26] **F. Molnar Jr.**, S. Sreenivasan, B. K. Szymanski & **G. Korniss**. Minimum dominating sets in scale-free network ensembles. *Scientific Reports* **3**, 1736 (2013). http://dx.doi.org/10.1038/srep01736.

[p37] N.T. Markov, M. Ercsey-Ravasz, D.C. Van Essen, K. Knoblauch, **Z. Toroczkai** & H. Kennedy. Cortical high-density counterstream architectures. *Science* **342**(6158), 1238406 (2013).

[p38] P. Singh, S. Sreenivasan, B. K. Szymanski & **G. Korniss**. Threshold-limited spreading in social networks with multiple initiators. *Scientific Reports* **3**, 2330 (2013). http://dx.doi.org/10.1038/srep02330

[p39] H. Aydinian, **É. Czabarka** & **L. A. Székely**. Mixed orthogonal arrays, k-dimensional M-part Sperner multi-families, and full multi-transversals. in *Information Theory, Combinatorics and Search Theory* (in memory of Rudolph Ahlswede) Lect. Notes. Comput. Sc. **7777**, 371-401, (2013), Springer-Verlag.

[p40] **P.L. Erdős**, Z. Király, **I. Miklós**. On the swap-distances of different realizations of a graphical degree sequence. *Combin. Prob. Comp.* **22**, 366-383, (2013).

[p41] **É. Czabarka**, M. Marsili and **L. A. Székely**. Threshold functions for distinct parts: revisiting Erdős-Lehner. in *Information Theory, Combinatorics and Search Theory* (in memory of Rudolph Ahlswede) Lect. Notes. Comput. Sc. **7777**, 463-471, (2013), Springer-Verlag.

[p42] **I. Miklós**, **P.L. Erdős** and L. Soukup. Towards random uniform sampling of bipartite graphs with given degree sequence. *Electron. J. Comb.* **20**(1) #P16, 1-51, (2013).

[p43] L. Lu, **A. Mohr** and **L. A. Székely**. Quest for negative dependency graphs. *Recent advances in harmonic analysis and applications*, Eds. D. Bilyk, L. de Carli, A. Stokolos, A. Pethukov, B. Wick. Springer Proceedings in Mathematics and Statistics, pp. 243-258. (2013).

[p44] Y. Dong, J. Tang, N. V. Chawla, How long will she call me? Distribution, social theory and duration prediction. In *Machine Learning and Knowledge Discovery in Databases*, pp. 16-31. Springer Berlin Heidelberg, (2013). http://dx.doi.org/10.1007/978-3-642-40991-2_2

[p45] L. Lu, **A. Mohr** and **L. A. Székely**. Connected Balanced Subgraphs in Random Regular Multigraphs Under the Configuration Model. *J. Comb. Math. & Comb. Comput.* (JCMCC) **86**, 111-123 (2013).

[p46] **Y. Dong**, J. Tang, S. Wu, J. Tian, **N.V. Chawla**, J. Rao, and H. Cao. Link Prediction and Recommendation across Heterogeneous Social Networks. *In Data Mining (ICDM), 2012 IEEE 12th International Conference on Data Mining*, pp. 181-190. IEEE, 2012. http://dx.doi.org/10.1109/ICDM.2012.140.

[p47] **Y. Yang**, **N.V. Chawla**, Y. Sun and J. Han Predicting Links in Multi-Relational and Heterogeneous Networks. *ICDM'12: The 12th IEEE Int. Conf. on Data Mining* Dec 10-13, Brussels, Belgium, pp 755-764 (2012). http://dx.doi.org/10.1109/ICDM.2012.144

Publications in preparation

[p48] S. Nagrecha, **N.V. Chawla**, H. Bunke. Recurrent subgraph prediction. (2015).

[p49] Y. Ren, D.C. Vural and **Z. Toroczkai**. Non-local effects from localized attacks in spatial networks without a cascade mechanism. (2015).

[p50] L. Lu, **A. Mohr**, **L.A. Székely**, Counting regular uniform hypergraphs using the Lovász Local Lemma. (2015).

[p51] **E. Czabarka**, **A. Dutle**, **F. Molnar**, **L.A. Szekely**, A local lemma algorithm for dominating sets. - technical report (2015).

[p52] **E. Czabarka**, **A. Dutle**, **I. Miklos**, Partition adjacency matrices. (2015).

[p53] **E. Czabarka**, K. Sadeghi, J. Rauh, T. Short, **L.A. Szekely**, On the number of non-zero elements in a Joint Degree Matrix. (2015).

[p54] **E. Czabarka**, **L.A. Szekely**, Research problems for the working seminar. - technical report, 2012

[p55] **L.A. Szekely**, S. Wagner, Hua Wang, Problems related to graph indices in trees, in: "Recent Trends in Combinatorics'', eds. A. Beveridge, J.R. Griggs, L. Hogben, G. Musiker, P. Tetali, Springer-Verlag. In press (2015).

[p56] **I. Miklos**, **H. Smit**h, Complexity results on the number of Single Cut and Join scenarios. (2015).

[p57] **H. Smith**, **L.A. Szekely**, Hua Wang, Shuai Yuan. On different "middle parts" of a tree. (2015).

[p58] **M. Varga**, D. Deritei, D. Barabasi, M Ercsey-Ravasz, **Sz. Horvat**, H. Kennedy, K. Knoblauch and **Z. Toroczkai**. Core-periphery and community structure in brain cortical networks. (2015).

[p59] Y. Ren and **Z. Toroczkai**. Weighted betweenness centrality approach to efficient network partitioning and community detection. (2015).

Theses, Dissertations (students graduated):

A. Strathman (Physics PhD, ND, 2013): "Applications of statistical mechanics to the modeling of social networks." Advisor: Z. Toroczkai

Y. Ren (Physics PhD, ND, 2015): "Betweenness Centrality and its Applications from Modeling Traffic Flows to Network Community Detection". Advisor: Z. Toroczkai

A. Mohr (Math PhD, USC, 2013): "Applications of the lopsided Lovász Local Lemma regarding hypergraphs". Advisor: L.A. Szekely.

H. Smith (Math PhD, USC, expected, Aug 2015): "Trees, Partitions, and other Combinatorial Structures". Advisor: L.A. Szekely.

C.D. Gaddy (Math MSc, USC, 2013): "Spectral analysis of randomly generated networks with prescribed degree sequences". Advisor: E. Czabarka

S. Trevino III (Physics PhD, UH, 2013) "Detecting Communities in Complex Unipartite and Bipartite Networks by Maximizing Modularity". Advisor: K.E. Bassler

A. Nyberg (Physics PhD, UH, 2014): "The Laplacian Spectra of Random Geometric Graphs." Advisor: K.E. Bassler

F. Molnar (Multidisciplinary Science Ph.D., RPI, 2014) "Computational analysis of complex systems: Applications to population dynamics and networks", Advisor: G. Korniss

Other Highlights

- N.V. Chawla received a 2012 IBM Watson Faculty Award and a 2013 IBM Big Data and Analytics Award.
- Z. Toroczkai was elected Fellow of the American Physical Society (APS), 2012.

Activities, Meetings, Interactions

- **Peter L. Erdős** of the Rényi Inst Math (RIM) visited Notre Dame Febr 24 – Mar 28, 2013. ND collaborated on two problems with him.
- **Tamara Kolda** (Sandia, and GRAPHS) visited Notre Dame , Apr 24-26. Talk: "Analysing and generating BIG networks". Discussed possible collaborations. Also hosted **Zoran Obradovic** (Temple U, GRAPHS), discussed collaborations.
- Several collaborative visits between Nitesh Chawla and **Zoran Obradovic** (Temple U, GRAPHS), discussed collaborations. Chawla and Obradovic are collaborating on a paper on health networks.
- **K.E. Bassler** (UH) and **Z. Toroczkai** (ND) were convener and co-conveners of an Advanced Study Group (ASG) at the Max Planck Inst. in Dresden Germany, on "Adaptive Networks". The Max Planck Inst hosted MAPCON12 – graph based modeling of complex systems conference. **P.L. Erdős** and **I. Miklós** visited **KEB** and **ZT** there and worked on the project.
- **Summer School in Network Science at USC**, May 20-31, 2013. Organizers: **É. Czabarka** and **L. Székely** (USC PIs) funded through USC. 40+ participants, talks by Fan Chung, Peter Mucha, Joel Spencer, van der Hofstad. GRAPHS participants **György Korniss (RPI)** and **Sonja Petrovic (IIT)** also presented, the meeting facilitated existing and new collaborations. http://imi.cas.sc.edu/events/summer-school-network-science/

- **É. Czabarka** and **L. Székely** visited the Rényi Inst. in June 2013 for collaborations on the project with **P.L. Erdős** and **I. Miklós**. They took PhD student Heather C. Smith to collaborate from USC funds.
- Several collaborative visits between **Sonja Petrovic**, **Despina Stasi** (IIT) and **É. Czabarka (USC)** and **Z. Toroczkai** (ND) at both IIT and ND in the Fall 2013.
- The whole team convened at the DARPA annual review meeting in Arlington, Nov 11-13 2013.
- **É. Czabarka** spent 3 months of her sabbatical (Aug 22-Nov 22, 2013) at Notre Dame, collaborating on the project.
- The PIs and visitors on this project gave numerous seminars and tutorials to the students and postdocs. Topics included network analysis methods, exponential random graphs, the probabilistic method (and the Lovász Local Lemma), graph sampling and Markov chains, etc.
- Featured Minisymposium: Data-Driven Modeling of Dynamical Processes in Spatially-Embedded Random Networks, *SIAM 2015 Conference on Applications of Dynamical Systems*, organized by **G. Korniss**, Snowbird, UT (May 17, 2015).
- Workshop on Statistical Physics of Social Networks, *International Conference on Statistical Physics (SigmaPhi2014)*, co-organized by **G. Korniss**, B.K. Szymanski, and C. Lim, Rhodes, Greece (July 7-11, 2014);

Presentations given

Unless specified otherwise, all presentations are invited talks (inverse chronological order).

1. **N. Derzsy**, X. Lin, A. Moussawi, B. Szymanski, **G. Korniss**, "Highly Damage-Resilient Dominating Sets in Complex Networks against Random and Targeted Attacks", *NetSci 2015 Meeting*, Zaragoza, Spain (June 4, 2015) Student presented poster.
2. **F. Molnár**, "Damage-resilient Dominating Sets in Real Complex Networks", Featured Minisymposium Data-Driven Modeling of Dynamical Processes in Spatially-Embedded Random Networks at the *SIAM 2015 Conference on Applications of Dynamical System*, Snowbird, UT (May 17, 2015). Invited student presentation.
3. K.E. Bassler, E. Frey, and R.K.P. Zia, "Diversity Driven Coexistence: Collective Stability in the Cyclic Competition of Three Species", *APS March 2015 Meeting*, San Antonio, TX, on March 5, 2015; abstract published in *Bulletin of the American Physical Society*, **60**(1); http://meetings.aps.org/link/BAPS.2015.MAR.S48.1
4. A. Nyberg and K.E. Bassler, "Eigenvalue Separation in the Laplacian Spectra of Random Geometric Graphs", *APS March 2015 Meeting*, San Antonio, TX, on March 4, 2015; abstract published in *Bulletin of the American Physical Society*, **60**(1); http://meetings.aps.org/link/BAPS.2015.MAR.Q44.10 .
5. Z. Toroczkai: *Opportunities for Nonlinear Sciences in the 21st Century*. Center for Nonlinear Studies (CNLS) Colloquium, Los Alamos National Laboratory, NM, Feb 10, 2015.
6. E. Czabarka: *Beyond degree sequences of graphs,* Institute for Mathematics and Its Applications Annual Seminar, Minneapolis, MN, Febr 2015.

7. H. Smith: Sampling Single Cut-or-Join Scenarios AMS Special Session on Network Science, 2015 Joint Mathematics Meetings, Henry B. Gonzalez Convention Center, San Antonio, TX, Jan 11, 2015
8. Sz. Horvát, *Reducing degeneracy in Exponential Random Graph models*, Northwestern Institute on Complex Systems, Evanston, IL, Jan 2015
9. Sz. Horvát, *Reducing degeneracy in Exponential Random Graph models*, Amaral Lab at Northwestern University, Evanston, IL, Jan 2015
10. E. Czabarka: *Beyond degree sequences of graphs*, Colloquium talk at Stellenbosch University, South Africa: Dec 17, 2014
11. L.A. Szekely, *Markov chains on Abelian groups give new constructions for the diamond problem*, Stellenbosch University, Stellenbosch, South Africa, Dec 2014
12. E. Czabarka: *Joint degree matrices and partition adjacency matrices* Fall Southeastern Sectional Meeting of the AMS, University of North Carolina at Greensboro , Nov 8 2014
13. Z. Toroczkai: *The Brain in Space and Time*. Journée scientifique Investissements d'Avenir de l'Université de Lyon, "La complexitée: quels défis pour demain?", Lyon, France, Nov 5, 2014. - Public lecture.
14. E. Czabarka: *Mixed orthogonal arrays and more-part Sperner families* Combinatorics Seminar at the Department of Mathematics, University of British Columbia, Vancouver, Canada, Oct 21 2014
15. L.A. Szekely, *Using the Lovasz Local Lemma as a tool for asymptotic enumeration*, University of British Columbia, Vancouver, Canada, Oct 2014
16. Z. Toroczkai: *A predictive model of the cortical network based on a distance rule*. European Conference on Complex Systems (ECCS'14). Workshop: "The Complex Brain", Institute for Advanced Studies, Lucca, Italy, Sep 24, 2014.
17. Z. Toroczkai: "Maximum entropy network models and applications." International Conference on Statistical Physics, Rhodes, Greece, Jul 7-14, 2014.
18. E. Czabarka: *Networks with similar assortativity – JDMs and PAMs* 2014 SIAM Conference on Discrete Mathematics, Minneapolis, June 16 2014
19. Z. Toroczkai: "Constrained graph construction problems in network modeling." SIAM Conference on Discrete Mathematics, Minneapolis, MN, Jun 16-19, 2014.
20. F. Molnár, N. Derzsy, E. Czabarka, L. Szekely, B. K. Szymanski, G. Korniss, "Scaling of Various Dominating Sets in Scale-Free Network Ensembles", NetSci 2014 Conference, Berkeley, CA (Jun 5, 2014) (poster);
21. F. Molnár Jr., N. Derzsy, B. K. Szymanski, G. Korniss, "Stability of Dominating Sets in Complex Networks against Random and Targeted Attacks", NetSci 2014 Conference, Berkeley, CA (Jun 5, 2014) (poster);
22. E. Czabarka: *Partition adjacency matrices* Combinatorics seminar at Dept. Math. University of Szeged, Hungary, Combinatorics Seminar, May 29 2014
23. L.A. Szekely, *Threshold functions for distinct parts: Erdos-Lehner revisited*, Szeged University, Szeged, Hungary, May 2014
24. Heather Smith: *Approximating the Number of Double Cut-and-Join Scenarios*. SIAM Student Chapter, University of South Carolina, Columbia, SC, Apr 30, 2014
25. F. Molnár, "Minimum Dominating Sets in Scale-Free Networks" at the IBM T. J. Watson Research, Yorktown Heights, NY (April 16, 2014).
26. Z. Toroczkai: *Big Data and the Brai*n. Central European University Roundtable on Network Science and Big Data. Budapest, Hungary, Mar 13, 2014.

27. F. Molnár, N. Derzsy, B.K. Szymanski, and G. Korniss, "Stability of Dominating Sets in Complex Networks against Random and Targeted Attacks", APS Mar 2014 Meeting, Denver, CO, on Mar 7, 2014; abstract published in Bulletin of the American Physical Society, 59(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.15.
28. N. Derzsy, F. Molnar, E. Czabarka, L. Szekely, B.K. Szymanski, and G. Korniss, "Scaling of Various Dominating Sets in Scale-Free and Empirical Networks", APS March 2014 Meeting, Denver, CO, on Mar 7, 2014; abstract published in Bulletin of the American Physical Society, 59(1); http://meetings.aps.org/link/BAPS.2014.MAR.Z17.10.
29. P. Karampourniotis, S. Sreenivasan, B.K. Szymanski, and G. Korniss, "Cascades in the Threshold Model with Multiple Initiators and Heterogeneous Threshold Values", APS March 2014Meeting, Denver, CO, on March 7, 2014; abstract published in Bulletin of the American Physical Society, 59(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.14.
30. K.E. Bassler and S. Hossein, "Symmetry in Critical Random Boolean Network Dynamics", *APS March 2014 Meeting*, Denver, CO, on March 7, 2014; abstract published in *Bulletin of the American Physical Society*, **59**(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.7 .
31. S. Hossein, F. Greil, and K.E. Bassler, "Critical Behavior in a Class of Heterogeneous Complex Systems", *APS March 2014 Meeting*, Denver, CO, on March 7, 2014; abstract published in *Bulletin of the American Physical Society*, **59**(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.8 .
32. A. Nyberg and K.E. Bassler, "Eigenvalue Separation in the Laplacian Spectra of Random Geometric Graphs", *APS March 2014 Meeting*, Denver, CO, on March 7, 2014; abstract published in *Bulletin of the American Physical Society*, **59**(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.9 .
33. F. Greil and K.E. Bassler, "Laplacian Spectra of Random Hyperbolic Geometric Graphs", *APS March 2014 Meeting*, Denver, CO, on March 7, 2014; abstract published in *Bulletin of the American Physical Society*, **59**(1); http://meetings.aps.org/link/BAPS.2014.MAR.Y16.10
34. C. Del Genio, K. Bassler, P. Erdos, I. Miklos, and Z. Toroczkai, "Sampling networks with prescribed degree correlations", *APS March 2014 Meeting*, Denver, CO, on March 7, 2014; abstract published in *Bulletin of the American Physical Society*, **59**(1); http://meetings.aps.org/link/BAPS.2014.MAR.Z17.13 .
35. E. Czabarka: *A gentle introduction to the Lovasz Local Lemma* Dept. Physics, University of Houston, Seminar of the Bassler Research Group, Feb 27, 2014
36. E. Czabarka: *Sampling graph ensembles with given assortativity* Dept. Physics, University of Houston, Seminar of the Bassler Research Group, Feb 24, 2014
37. L.A. Szekely, *Counting graphs with the Lovasz Local Lemma*, Joint Mathematics Meeting, Extremal and Structural Graph Theory Session, Baltimore, MD, Jan 2014
38. E. Czabarka: *Sperner type problems and design of experiments* Dept. of Math, Zhejiang University, Hangzhou, China, Dec 19, 2013
39. E. Czabarka: *Sperner type problems and design of experiments* Dept. of Math. Tongji University, Shanghai, China, Dec 16 2013
40. E. Czabarka: *Partition adjacency matrices*, Dept. of Math, Nanjing Normal University, Nanjing, China, Dec 13, 2013
41. L. Szekely, *Using the Lovasz Local Lemma in asymptotic enumeration*, Nanjing Normal University, Nanjing, P.R. China, Dec 2013
42. E. Czabarka: *Partition adjacency matrices* Dept. of Math., Tongji University, Shanghai, China, Dec 9 2013

43. L.A. Szekely, *Using the Lovasz Local Lemma in asymptotic enumeration*, Tongji University, Shanghai, P.R. China, Dec 2013
44. Z. Toroczkai: *A Distance Rule Based Predictive Model of the Cortical Brain Network.* Network Frontier Workshop, Northwestern University, Dec 4-6, 2013.
45. E. Czabarka: *Partition adjacency matrices* Dept. of Math, University of Louisville, KY, Oct 23, 2013
46. E. Czabarka: *A gentle introduction to the Lovasz Local Lemma and its applications I, II* University of Notre Dame, Interdisciplinary Center for Network Science & Applications, (2+2 hours), Oct 8,10, 2013
47. E. Czabarka: *Connecting Sperner problems to mixed orthogonal arrays*, University of Notre Dame Combinatorics Seminar, Oct 7, 2013
48. E. Czabarka: *On realizations of a joint degree matrix* Fall Southeastern Sectional Meeting of the AMS, University of Louisville, KY, Oct 5 2013
49. L.A. Szekely, *Extremal values of ratios: distances vs. the number of subtrees*, 1092 AMS Sectional Meeting, Extremal Combinatorics session, Louisville, KY, Oct 2013
50. L.A. Szekely, *Using the Lovasz Local Lemma in asymptotic enumeration*, Monash University, Melbourne, Australia, Sep 2013
51. A. Dutle: *Graph Theory in the Information Age* Department Colloquim, Northern Kentucky University, Sep 13, 2013.
52. E. Czabarka: *From Sperner-type problems to mixed orthogonal arrays* Colloquium, Dept. of Applied Math., Illinois Institute of Technology, Aug 26, 2013
53. A. Dutle: *Realizations of Joint Degree Matrices* Second PRIMA Congress, Shanghai, China, Jun 25, 2013.
54. Sz. Horvát, *A gentle introduction to the statistical mechanics of networks and exponential random graph models*, Frankfurt Institute for Advanced Studies, Frankfurt, Germany, Jun 2013.
55. F. Molnar, "Minimum Dominating Sets in Scale-Free Network Ensembles", NetSci 2013 Conference, Copenhagen, Denmark (June 6, 2013).
56. Sz. Horvát, *Introduction to Exponential Random Graph models*, Max Planck Institute for the Physics of Complex Systems, Dresden, Germany, Jun 2013
57. Z. Toroczkai: *Functional modularity from simultaneous adaptation to multiple constraints.* NetSci 2013. The International School and Conference on Network Science. Workshop "Network Models in Cellular Regulation", Copenhagen, Denmark Jun 4, (2013).
58. F. Molnár, "Minimum Dominating Sets in Scale-Free Network Ensembles", Summer School on Network Science, University of South Carolina, Columbia, SC (May 27, 2013).
59. A. Dutle: *Degree Sequence and Joint Degree Matrix Models* Summer School on Network Science, University of South Carolina, May 22, 2013.
60. Z. Toroczkai: *Predicting Traffic Changes in the Wake of Geo-Localized Damages in Large-Scale Transportation Networks*. SIAM Conference on Applications of Dynamical Systems, May 19-23, Snowbird, Utah, USA.
61. L.A. Szekely, *Threshold functions for distinct parts: revisiting Erdos-Lehner*, SIAM-SEAS meeting, Extremal Combinatorics session, University of Tennessee-Knoxville and Oak Ridge National laboratory, TN, Mar 2013
62. A. Dutle: *Realizing Joint Degree Matrices* Extremal combinatorics special session, SIAM-SEAS 2013 annual meeting, University of Tennessee Knoxville, Mar 23, 2013.

63. F. Molnár, "Scaling of Minimum Dominating Sets in Various Scale-Free Network Ensembles", APS March Meeting 2013, Baltimore, MD (Mar 22, 2013).
64. F. Molnár, "Scaling of Minimum Dominating Sets in Various Scale-Free Network Ensembles" 4th Workshop on Complex Networks (CompleNet 2013), Berlin, Germany (Mar 14, 2013).
65. L.A. Szekely, *M-part Sperner multifamilies, multitransversals, and mixed orthogonal arrays* (30 min) Carolina Math Seminar, Citadel, Charleston SC, Oct 2012
66. L.A. Szekely, *M-part Sperner multifamilies, multitransversals, and mixed orthogonal arrays* (30 min), International Conference on Advances in Interdisciplinary Statistics and Combinatorics, Combinatorics Session Oct 5, 2012, Greensboro, NC
67. E. Czabarka: *Mixed orthogonal arrays, k-dimensional M-part Sperner multi-families and full multi-transversals* Advances in Interdisciplinary Statistics and Combinatorics, University of North Carolina Greensboro, Oct 5 2012
68. L.A. Szekely, "Constructions for the diamond problem", Search Methodologies III Conference, ZiF, University of Bielefeld, Germany, Sep 2012
69. E. Czabarka: *Mixed orthogonal arrays, k-dimensional M-part Sperner multi-families and full multi-transversals* Search Methodologies III, Zentrum fur Interdisziplinare Forschung, Sep 1, 2012

Technical Accomplishments

T1. Constrained Graph Ensembles

Complex networks of national importance such as social networks and infrastructure networks are dynamical structures containing both structural and temporal patterns of interest (such as instabilities, failure signatures, adversarial dynamics) the extraction of which is critical for the defense of these systems. However, these signatures are embedded in a stochastic environment with a very large phase space, presenting formidable challenges for the development of tools that would analyze these systems and extract these patterns. Other, compounding factors include lack of information and the dynamic nature of the networks. The interactions within these networks are mediated and determined by physical processes and thus they obey natural physical constraints such as finite rate of communication, finite throughput and transport capacity, limited information storage ability, etc., constraints that have imprints on the possible graph structures that can form such networks. However, as any physical theory, a relevant understanding and modeling of networks must be rooted in and developed from data. Data not only provides first-hand information, but also serves as a set of constraints on network models. As data is hardly ever complete and is often noisy, there can be a large number of possible network structures obeying the given data. The set of all graphs obeying the empirical data forms *a graph ensemble*. A central goal of this proposal is to develop a systematic mathematical approach towards modeling complex networks using constrained graph ensembles. Constraints (data) can be obeyed either A) verbatim (sharp constraints) or B) on average (average constraints). An example for case A) is the set of all graphs with the same given degree sequence. In case B) the constraints are obeyed by ensemble averages, such as by the set of all graphs with a given expected degree sequence. For a successful data driven modeling, both the sharp constraints and the average constraints cases need to be studied. Once we developed an understanding on how the constraints determine the properties of the corresponding graph ensemble,

it enables the building of probabilistic models for questions of interest related to this ensemble, for e.g., "How do we determine efficiently and with good approximation the smallest set of nodes and edges that when influenced/controlled will have the largest effect on the whole network? Can we predict better than by pure chance the appearance of certain links or subgraphs in the network? What are the structural and dynamical vulnerabilities of a network? In case of a localized attack or breakdown on a large network, how can we predict the new patterns of flow/dynamics that the network settles in and its new vulnerabilities? , etc.

T1.1 Sharp Constraints

Existence, Construction, Sampling and Counting Problems of graphs with given constraints

A graph obeys a given *constraint* (a well defined graph measure whose value is given by data, or assumed known) *sharply* if the corresponding measure on this graph has the same value as in the data. A *sharply constrained ensemble* is the set of all graphs that obey the given constraint (or constraints) sharply. Let $\boldsymbol{x} = (x_1, \dots, x_K)$ be a set of graph measures acting as constraints, for example the list of degrees of the nodes, or the number of connections between given degrees for all degree pairs (degree correlations), the number of triangles incident on a specific nodes, etc. For a given graph G, the value of that measure on this graph is $x_j(G)$. A graph ensemble on N nodes, constrained sharply by $\boldsymbol{x}$ is: $\mathrm{G}_1(\boldsymbol{x}) = \{G \in \mathrm{G}_1 | x_j(G) = x_j, \forall j \in \{1, \dots, K\}\}$

where G_1 denotes the set of all graphs on N nodes and x_j is the value of the constraint for that measure fixed by the data. When modeling real-world networks as constrained by existing data, we typically have to deal with the following four fundamental problems:

Existence. Under what conditions on $\boldsymbol{x}$, $\mathrm{G}_1(\boldsymbol{x}) \neq \emptyset$? – i.e., are there any graphs satisfying the given constraints?

Construction. How to build any (or all) member(s) of $\mathrm{G}_1(\boldsymbol{x})$?

Sampling. How to sample by some distribution (typically uniformly) members of $\mathrm{G}_1(\boldsymbol{x})$?

Counting. How to compute or estimate $|\mathrm{G}_1(\boldsymbol{x})|$? ($|\mathrm{G}_1(\boldsymbol{x})| \leq |\mathrm{G}_1| 2^{-N^2}$)

During this project we made significant contributions to all the 4 types of fundamental problems above for two large classes of constraints: a) degree sequence based and b) joint degree matrix based.

a) For degree sequence based constraints we have $\boldsymbol{x} = \boldsymbol{d} = (d_1, \dots, d_N)$ for undirected graphs, where d_i is the degree of node i; $\boldsymbol{x} = \boldsymbol{dd} = ((d_1^{in}, d_1^{out}), \dots, (d_N^{in}, d_N^{out}))$, for directed graphs, where $d_i^{\pm}$ are the in- and out- degrees respectively; $\boldsymbol{x} = \boldsymbol{bd} = (\boldsymbol{d}|_U, \boldsymbol{d}|_W)$, for bipartite graphs where U and W are the sets of nodes on the two sides of the partition and $U \cup W = V$, and $\boldsymbol{d}|_{U(W)}$ are the degree sequences of the nodes within their partitions.

b) A joint degree matrix (JDM) specifies the number of edges between nodes of given degrees for all degree pairs. This is a stronger constraint than just degree sequence based (given in a) above) as it determines both the degree sequence and the degree-degree correlations. If we partition the nodes into groups such that all nodes within a group have the same degree, i.e., $V_\alpha = \{v \in V | d(v) = \alpha\}$, then $\boldsymbol{x} = \boldsymbol{J} = [J_{\alpha\beta}]_{\alpha,\beta \in d(V)}$ where $J_{\alpha\beta} = |\{(v,u) \in E \; G(v) : v \in V_\alpha, u \in V_\beta\}|$, i.e., it specifies the number of edges between nodes of degrees α and β.

There are many applications for both degree-based constraints and joint degree matrix based constraints. For example, the latter can be used to model social networks with prescribed degree-correlations, thus modeling the assortative mixing property found to be a strong characteristic of these networks.

In the following we briefly summarize the results obtained related to the four fundamental problems during the project for the two classes of constraints. For a), the *Existence* problem has already been solved expressed as the Erdos-Gallai and the Havel-Hakimi theorems. For b), i.e., JDMs, we have provided a clean and short proof in Ref **[p24]** to the corresponding existence theorem.

Graph *Construction* follows two main approaches: 1) Direct construction, where we start from an empty graph and we add/create edges such that the constraints are respected. The key issues here are to make sure that any graph from the ensemble $G_!(\boldsymbol{x})$ can be built by the construction algorithm and that it is done efficiently (poly time). 2) Via swap/switch based methods where we start from a graphical realization $G \in G_!(\boldsymbol{x})$ then we move edges around in the graph, usually by some swaps/switches in a way to preserve graphicality and also the constraints, arriving at another member $G' \in G_!(\boldsymbol{x})$. The key issue is to make sure that all members of $G_!(\boldsymbol{x})$ can be reached by the proposed swap operation.

1) Prior to this project we have solved the direct construction problem for degree-based constraints in Ref [Kim2009] for undirected graphs and Ref [Erdos2010] for directed graphs. To create the corresponding algorithms we had to solve another existence problem, namely to prove the necessary and sufficient conditions for graphicality of degree sequences that are restricted with forbidden edges forming a *k*-star on an arbitrary node *i*. In other words, these conditions guarantee the existence of (or show that there aren't such) graphs that realize the degree sequence such that a set of edges is forbidden to appear (in this case a star of *k* edges). This turned out to be a fundamental theorem for all direct construction problems, including for joint-degree matrix based constraints. We have recently also solved the direct construction problem for JDM based constraints, with the algorithm described in Ref **[p9]**.

2) Swaps/switches based graph construction has been the standard method before our direct construction work for degree-based constraints. In this case swapping two ends of two independent edges ("2-swaps") leads to a new graph with the same degree sequence. Ryser and Taylor have shown that this operation when repeated is able to visit all members of $G_!(\boldsymbol{d})$. We start from a graphical realization (for example generated by the Havel-Hakimi algorithm) then keep doing these 2-swaps. For JDM based constraints we introduced the so-called restricted 2-swap operation (RSO) that allows only swaps between two edges that have at least one end in the same degree class. We have proven in Ref **[p24]** that an RSO based Markov Chain Monte Carlo (MCMC) algorithm will preserve the JDM and visits all members of the ensemble $G_!(\mathbf{J})$.

Graph *Sampling* also follows two approaches since both graph construction methods (direct and swap based) can be used to generate graphs at random from $G_!(\boldsymbol{x})$. Here the goal is to sample graphs uniformly from $G_!(\boldsymbol{x})$.

1) Using the direct construction approach, for degree-based constraints we solved this sampling problem earlier. Since the direct construction method does not generate the graphs uniformly at random, we need to reweight the samples accordingly, to count for this bias. However, our algorithm does compute the sample weights as well, so this can be done. We demonstrated this in [Genio2010] for undirected graphs and in [Kim2012] for directed ones. Expanding on our results

from [Genio2010] and [Kim2012] we have recently provided a similar direct construction based uniform sampling algorithm in **[p9]** for JDM based constraints.

2) Can we generate efficiently and uniformly at random graph samples using swap/switch based MCMC algorithms? For degree-based constraints this question has been around since it was proposed by Kannan, Tetali and Vempala in [Kannan1999]. The main issue here is the so-called mixing-time problem: How long do we need to keep running the swap MCMC algorithm to guarantee quasi-independent graph samples? In particular, the conjecture is that the simple swap-based MCMC for degree-sequence based constraints is fast mixing, namely, the relaxation time $\tau_{mix} = (1 - \lambda^*)^{-1}$ of the Markov Chain has an upper bound that grows only polynomially with the number of nodes N. Here $\lambda^* = \max\{|\lambda_2|, |\lambda_N|\}$ where $1 = \lambda_1 > \lambda_2 \geq \ldots \geq \lambda_N \geq -1$ are the eigenvalues of the row-stochastic transition matrix $\boldsymbol{P}$ of the Markov Chain. This conjecture is notoriously difficult to prove (or disprove) and it is still an open problem. Since MCMC based sampling methods are the industry standard in modeling stochastic systems, this is also an important problem. There have been results obtained for specific degree sequences: [Kannan1999] have shown fast mixing for regular bipartite graphs, [Cooper2007] has shown it for regular undirected graphs and [Greenhill2011] has shown it for regular directed graphs (these are special degree sequences). We have obtained major results during this project on this problem: In **[p42]** we have proven the conjecture for half-regular bipartite graphs, i.e., bipartite graphs for which only on one side of the bipartition the nodes have to be of same degree, with the other side having an arbitrary degree sequence. This is a very technical proof on 50 pages. The next natural question was whether we could sample uniformly and in poly-time using this MCMC swap algorithm such as to avoid certain subgraphs. This question was answered affirmatively for half-regular bipartite degree sequences avoiding a k-star sitting on an arbitrary node and a 1-factor (which is a perfect matching between the two node classes). These results have been presented in **[p12]**. For JDM based constraints using MCMC sampling the same question seems even harder to answer. However, we recently have been able to make a significant advance, published in **[p18]**. We were able to provide an MCMC sampling algorithm that generates graph samples uniformly and in poly time from the subset of balanced graphical realizations within $G_!(\mathbf{J})$. A graphical realization is balanced if the degrees within all degree classes are as uniformly distributed as possible. Every graphical JDM admits balanced realizations. The next step will be to try to remove this constraint that was needed for our proof techniques.

Counting problems are the hardest among the four problem classes in graph ensemble based modeling of networks. The goal here is to compute or estimate the size of the ensemble, i.e., $|G_!(\boldsymbol{x})|$. This is an important question in network modeling because the size of the ensemble indicates how constraining is the set of given measures $\boldsymbol{x} = (x_!, \ldots, x_!)$. We obtained two major results on counting problems, one that uses the Lovasz Local Lemma (described later) and another that we describe as follows. It has been shown that in terms of computational complexity, counting problems are harder than uniform sampling ones, which in turn are harder that construction problems which in turn are harder than existence problems. Typically, counting problems are in #P complexity class, which means that they are even harder in general then NP-hard problems. We were able to provide one result to this difficult class of problems in **[p12]**. In particular we have shown that the degree-based problem for half-regular bipartite sequences that excludes a 1-factor and an arbitrary k-star is self-reducible [Jerrum1986] and thus, based on a theorem from [Jerrum1986], it implies that there is a Fully Polynomial Randomized Approximation Scheme (FPRAS) (see [Jerrum1986] and [Vazirani2003]), i.e., an algorithm that estimates $|G_!(\boldsymbol{x})|$ in poly

time. This also implies that there is also a Fully Polynomial Almost Uniform Sampler (FPAUS), i.e., an algorithm that generates graph samples with these constraints almost uniformly, in polynomial time.

Graph similarity measures

Blair Sullivan's work [Adcock2013], [Adcock2014] on understanding the intermediate-scale network structure, led to the hypothesis that these structures are tree-like. In this direction L.A. Székely with former student H. Wang studied the analogy of the sum of distances, called the Wiener index, and the number of subtrees in a tree. They concluded that in extreme cases their behavior is just the opposite [Szekely2005], [Szekely2006]. Wagner [Wagner07] made an analysis of correlation the between a number of pairs of tree indices, and found the highest (negative) correlation between the Wiener index and the number of subtrees.

The Wiener index was introduced by the chemist H. Wiener in 1947. The Wiener index is proportional to the boiling temperatures of alkanes, and is perhaps the most frequently used topological index in mathematical chemistry. (Topological) indices were introduced to differentiate molecules/graphs/networks. Ideally, a topological index should measure the similarity of two networks: the further two networks are in structure, the higher difference in value. A "perfect" and computable topological index would solve the graph isomorphism problem, an unlikely outcome. Graph similarity measures studied so far have been based on distances or degrees. For example, sophisticated network entropy measures are also based on distances [Dehmer2008], [Dehmer2011].

Recall that the center is the set of vertices that minimize the largest distance observed in a vertex, the centroid is the set of vertices that minimize the sum of distances from a vertex to all other vertices, and the subtree core the set of vertices that are included in the largest number of subtrees. It is known that each of the center, centroid, and subtree core consists of one or two adjacent vertices. An ongoing work **[p57]** how far these different middle parts can be from each other in a tree.

In recent work **[p35]**, **[p28]**, H. Wang and L.A. Székely extended the analogy between several concepts involving distances and number of subtrees in trees to certain ratios of distances vs. ratios of number of subtrees. While the actual extreme values of these ratios do not agree, the corresponding extremal trees tend to be very similar, sometimes the same, as in [Barefoot1997], where extremal values for certain ratios of distances were investigated, pointing to a deeper connection between the two kinds of similarity measures. **[p11]** solves the extremal problems for corresponding ratios for eccentricity. L.A. Székely has been invited to write a survey paper for the volume "Trends in Combinatorics", which is a conclusion of the Special Year in Discrete Mathematics at the Institute for Mathematics and its Applications in Minneapolis, his contribution **[p55]** is on graph indices on trees, covering the topics above.

Complexity results for the single cut and join model

Palmer [Palmer1988] compared the mitochondrial genomes of cabbage and turnip, which are very closely related. To their surprise, these genomes, which are almost identical in gene content, differ dramatically in gene order. This and many other studies convincingly proved that genome rearrangements represent a common mode of molecular evolution. Gene reordering is particularly fast in cancer. "Double-Cut-and-Join" (DCJ) and "Single-Cut-or-Join" (SCJ) are the two basic combinatorial models for genome rearrangement, which allow rigorous proofs. Understanding these models is a key towards advanced algorithms for phylogeny reconstruction based on gene order data. Given a collection of observed taxa related by a phylogenetic tree, we are interested in sampling from the small parsimony scenarios i.e. ancestors together with the steps in which gene re-ordering

occurred. For DCJ, a Fully Polynomial time Randomized Approximation Scheme (FPRAS) and a Fully Polynomial Almost Uniform Sampler (FPAUS) are available for counting and sampling most parsimonious rearrangement scenarios between two genomes [Miklos2012]. [Miklos2014] show that both the sampling and counting problems are easy for two genomes in the SCJ model, however, for an arbitrary number of genomes related by a binary phylogenetic tree, the counting and sampling problems become hard, i.e. FPRAS or FPAUS exists for the most parsimonious SCJ scenario, then RP = NP. [Miklos2014] poses an open problem about the analogous problems in the SCJ model on star trees.

Ph.D. student H. Smith and I. Miklos first tried to prove an analogue of [Miklos2012] in this case, but the expected sampling analogue of a lemma failed. Then they proved #P-completeness for computing the total weight of all most parsimonious phylogenetic histories, no FPRAS unless RP = NP result for sampling most parsimonious ancestor labelings with various weighting functions, and also an NP-hardness result to compute the minimum of other weighting functions on labelings **[p56]**. These results, however, do not answer the open problem of [Miklos2014], but make steps towards it.

Symmetry in Heterogeneous Complex Systems

Using Boolean networks as prototypical examples, we have examined the role of symmetry in the dynamics of heterogeneous complex systems. Complex systems often differ from more traditional condensed matter systems because they consist of heterogeneous components, which makes their analysis and achieving any sort of general understanding of their behavior difficult. A question that naturally arises then is: Can symmetry in heterogeneous complex systems be exploited to simplify their analysis and obtain fundamental insights into their dynamics? In order to answer this question, it is essential to determine the role that symmetry has in controlling the behavior of these systems, and what it can tell us about their structure. It is also important to know whether symmetry can be used to distinguish systems with different dynamics. Boolean networks, which were first introduced as models of gene regulatory networks, are an ideal type of complex system with which to answer these questions. This is because Boolean networks, despite being relatively simple, have essential features of complex systems, including heterogeneous structure and dynamics of their constituent parts, and display nontrivial dynamics. Most notably, there exists a continuous phase transition in their dynamical behavior.

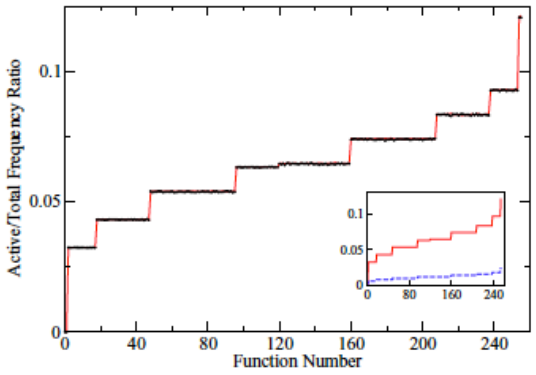


Figure 2. Comparison of the analytical calculations and simulation results for frequency at which nodes with a given Boolean function are active in the dynamics of networks consisting of nodes with in-degree K = 3. In the main figure, the black circles show simulation results, and the red line shows the results of analytical calculations. The inset compares the analytical results with (red solid line) and without (blue dashed line) including the effect of fluctuations. Note the necessity of including the effects of fluctuations in order to accurately predict the networks' behavior.

We have shown that the symmetry of the dynamics of Boolean networks, especially in critical states, is a controlling feature that can be used both to greatly simplify analysis and to characterize different types of dynamics. Symmetry in Boolean networks is found by determining the frequency at which the various Boolean output functions occur. In Boolean networks some nodes have frozen dynamics, while others are active. Using both

simulations and analytic calculations that include the renormalizing effects of stochastic fluctuations, we have determined the frequency that nodes that are active have a particular Boolean function. As can be seen in Fig. 2, there are classes of functions that consist of Boolean functions that behave similarly. These classes are orbits of the controlling symmetry group.

The symmetry found is one that preserves canalization, a form of network robustness. For networks with K=2 inputs per node, we have found a minimal canalization preserving symmetry group. It can be presented in terms of 3 generators (plus identity). A Cayley diagram showing the effects of group operations is shown in Fig. 3. We have also compared it to a different symmetry known to control the dynamics of an evolutionary process that allows Boolean networks to organize into a critical state. Our results demonstrate the usefulness and power of using the symmetry of the behavior of the nodes to characterize complex network dynamics, and introduce an alternative approach to the analysis of heterogeneous complex systems. These results have been published in **[p32]**.

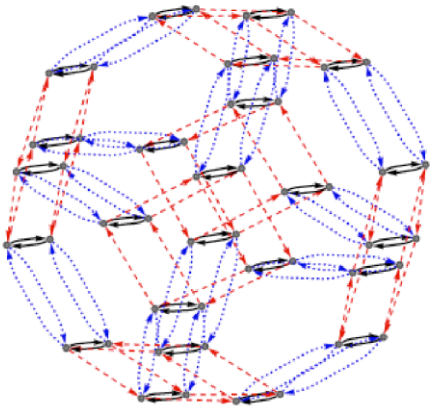


Figure 3. Cayley diagram of the minimal canalization preserving symmetry group for Boolean functions with K = 2 inputs. The elements of the group are the nodes of the graph, shown as gray dots. Directed edges of the graph indicate the effect of combining one of the group generators with an element. Arrows with different colors correspond to the effect of the three different generators of the group; rotation R, pruning N, and parity P operations are shown as red dashed, blue dotted, and black solid arrows, respectively.

T1.2 Average Constraints

Solving the degeneracy problem of exponential random graphs (ERGs)

Our understanding and modeling of complex systems and complex networks is always based on partial information. The only *general* principled way of creating predictive models that incorporate data from observations (the incomplete information that is known about the system) is using the principle of entropy maximization, as described by E. T. Jaynes [Jaynes1957]. By applying the maximum entropy principle to modeling networks, we obtain a family of models called Exponential Random Graphs.

Exponential Random Graph (ERG) models are widely used for drawing inferences from network data [Robins2007]. They are considered the only *general* modeling approach that works directly with networks rather than simply network properties. Unfortunately, their practical applicability is limited by the so-called *degeneracy problem*, as will be explained in more detail below [Handcock2003]. We have developed a general and practical approach for avoiding this problem and significantly expand the applicability of ERG models **[p8].** The results go beyond network models and have direct uses for general maximum entropy based modeling.

Exponential Random Graphs

ERG models are most convenient to understand through the problem of *sampling random graphs with constraints*. In many cases we only have partial information about a network, and may only know certain network properties instead of all the connections. For example, in a survey of the web of sexual connections [Liljeros2001] people were asked about how many sexual partners they

had, but not who these partners were. In other words, the degree sequence of the network is known, but not the actual connections. The exact same degree sequence can correspond to many different networks, so the problem of sampling from these choices presents itself.

From now on we will refer to a quantity m characterizing a network as a *graph measure*. Examples of graph measures include the number of edges in the network, the number of certain types of subgraphs within the network (e.g. the triangle count), the degree of a particular vertex, or even more complex measures such as assortativity. These can be thought of as functions $m: \mathcal{G} \to \mathbb{R}$ where $\mathcal{G}$ denotes the set of all graphs. A *sharp constraint* fixes the value of one or more graph measures precisely. For example, we may ask to uniformly sample graphs having precisely $m = m^*$ edges. Working with sharp constraints is typically difficult as there is no general approach that works for an arbitrary choice of constrained graph measures.

Alternatively we may use an *average constraint*, i.e. assign a probability $P(G)$ to each graph G so that the average value of the measure m over the set $\mathcal{G}$, $\langle m \rangle = \sum_{G \in \mathcal{G}} m(G) P(G)$, is equal to m^*. Such a constraint doesn't fully determine $P(G)$ as many different distributions will have the same average. Which choice for $P(G)$ is going to be the best one? The principle of maximum entropy states that the distribution having maximal information entropy $S = -\sum_{G \in \mathcal{G}} P(G) \ln P(G)$, while still satisfying the constraint, is going to be the least biased one. This distribution describes our limited knowledge about the system best. The graph ensemble obtained using an average constraint is called an Exponential Random Graph model.

Unlike in the case of sharp constraints, there is a straightforward method to construct a graph ensemble constrained in average. Carrying out the entropy maximization using the method of Lagrange multipliers yields the distribution

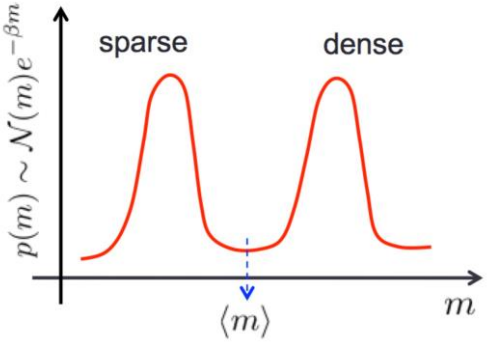
$$P(G) = \frac{1}{Z(\beta)} e^{-\beta m(G)} \quad \text{where} \quad Z(\beta) = \sum_{G \in \mathcal{G}} e^{-\beta m(G)}$$

where β is the Lagrange multiplier parametrizing the distribution. The average $\langle m \rangle$ will depend on parameter β: $\langle m \rangle = \langle m \rangle_\beta$. Therefore, to complete the task of constructing the distribution we must find the parameter $\beta = \beta^*$ that satisfies $\langle m \rangle_{\beta^*} = m^*$. This is called *fitting the model to the data* m^*. This is typically a difficult task that needs to be carried out numerically: finding $\langle m \rangle_\beta$ for a given parameter β is usually done using Markov Chain Montre Carlo sampling.

ERG models may be based on more than a single graph measure. Let us denote an ERG model based on graph measures $m_1, m_2, \ldots, m_k$ by $\mathrm{ERG}(m_1, m_2, \ldots, m_k)$.

The degeneracy problem

Degeneracy is known problem that limits the practical applicability of ERG models. To understand degeneracy, we first need to understand how ERG models are used in practice.



with none realizing the average $\langle m \rangle$.

ERG models are typically used as follows: there is some empirical data which is a partial or full observation of a network. This network is assumed to be a typical representation of its class. For example, if the measurement is the connectivity (edge density) of a high school friendship network, it is reasonable to assume that other high schools of similar size will have similar connectivities. Based on this assumption, various properties of the network (graph

measures) can be used as average constraints to construct an ERG model that describes this class of networks and can be used to predict other graph measures as well.

Unfortunately, depending on the choice of constrained graph measures it may happen that none of the graphs sampled from the ERG model have properties close to the input constraint. For example, all graphs may be either sparse or dense, while the average edge count is inbetween (see Fig. 4). In other words, the constraint $\langle m \rangle = m^{!}$ is no longer a typical value of the ensemble and predictions fail. This situation is called degeneracy [Handcock2003] and it has been one of the biggest obstacles in the applicability of ERG models.

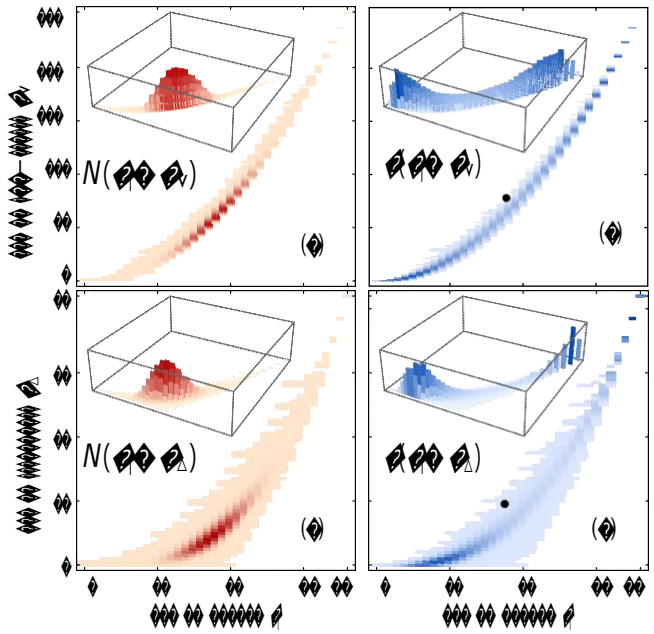
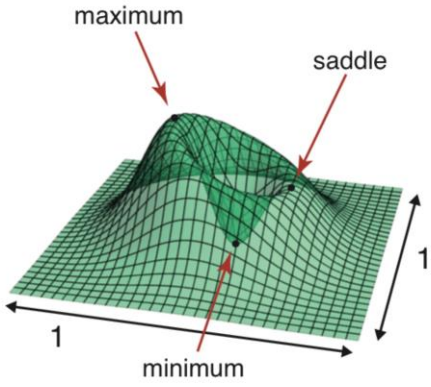


Figure 5. Examples of degenerate ERG models. (a) $\mathcal{N}(m_{|}, m_{\vee})$ obtained using exact enumeration of all graphs on 9 vertices. Darker areas show larger graph counts. The white areas do not correspond to any graph. (b) $p(m_{|}, m_{\vee}; \beta_{\vee} = -0.313, \beta_{|} = 2.20)$ (c) $\mathcal{N}(m_{|}, m_{\Delta})$ for graphs on 9 vertices (d) $p(m_{|}, m_{\Delta}; \beta_{\Delta} = -0.610, \beta_{|} = 1.24)$.

Let $\mathcal{N}(m_{|}, m_{|}, \ldots)$ denote the number of graphs having properties $m_{|}, m_{|}, \ldots$. The probability to draw a graph with property m from $\mathrm{ERG}(m)$ is $p(m; \beta) = \frac{!}{!(!)} \mathcal{N}(m) e^{!!"}$.

We say that a distribution $p(m)$ is degenerate if it is not concentrated around its average. We say that an ERG model is degenerate if for some value of β the distribution $p(m; \beta)$ becomes degenerate. This *typically* happens when $p(m)$ is not unimodal. Two examples of the degenerate models are shown in Fig. 5: $\mathrm{ERG}(m_{|}, m_{\vee})$ and $\mathrm{ERG}(m_{|}, m_{\Delta})$, where $m_{|}$ denotes the edge-count of a graph, $m_{\vee}$ denotes the number of two-star subgraphs (a two-star has a central node connected to two outer nodes) and m_{Δ} denotes the number of triangle subgraphs. These two models are often used to illustrate degeneracy [Park2004,Park2005] and we will continue to use them below.



Currently, the state of the art solution to the degeneracy problem is to limit the types of graph measures that are used to construct ERG models to "known good" ones that do not lead to degeneracy [Snijders2006]. Unfortunately the known good measures that were proposed in the literature do not have simple intuitive interpretations. In practical modeling one would prefer to let the problem at hand dictate the constraints instead. For example, when modeling social networks, choosing the triangle count as one of the $\beta_{|}$ a direct owever,

Figure 6. A "skewed hat" distribution is an example of a degenerate distribution that has a single local maximum. However, it also has a saddle point and a minimum.

and the triangle count simultaneously is known to lead to degeneracy. When working with food webs, ecologically relevant measures, such as the number of trophic levels, would be appropriate. Below we propose a practical solution for avoiding degeneracy that works regardless of the choice of graph measures, thus significantly extending the applicability of ERG models. The biggest advantage of our method is that it gives freedom in choosing the most appropriate constraints.

Why does degeneracy happen?

We can formalize the concept of degeneracy as follows: a distribution is degenerate if all its stationary points are local maxima. Note that when multiple graph measures are used, requiring unimodality is not sufficient: Figure 6 shows an example of a two-dimensional distribution with a single maximum wich is not concentrated around its average. However, this distribution also has a minimum and a saddle point.

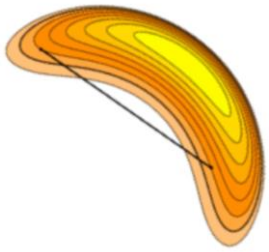


Figure 7. Example showing that if the contour lines of a function are not geometrically convex then the function is not concave.

Theorem: Let $\boldsymbol{m} = (m_1, m_2 \dots m_k)$ be the vector of constrained measures. The distribution $\mathcal{N}(\boldsymbol{m})$ is non-degenerate if the function $\mathcal{N}(\boldsymbol{m})$ is log-concave, i.e. if $\mathcal{N}(\boldsymbol{m})$ is a concave function.

This implies that if the model ERG($\boldsymbol{m}$) is non-degenerate then all contour lines of $\mathcal{N}(\boldsymbol{m})$ (i.e. $\mathcal{N}(\boldsymbol{m}) = \text{const.}$ curves) are convex (see Fig. 7).

Notice in Fig. 7 that both $\mathcal{N}(m_1, m_2)$ and $\mathcal{N}(m_1, m_\Delta)$ also have contour lines which are not convex. This explains why both ERG(m_1, m_2) and ERG(m_1, m_Δ) are degenerate. In both of these cases this geometric concavity results from the fact that the two constrained measures are not independent. For example, a graph with many edges (m_1) *must* also have many two-star subgraphs (m_2).

Typically, degeneracy is a result of such interdependencies between the constrained measures, and degeneracy usually appears in practical applications when more than a single graph measure is constrained. An intuitive way to see how degeneracy arises when the domain of $\mathcal{N}(\boldsymbol{m})$ is not convex is to notice that the average $\langle \boldsymbol{m} \rangle$ can take values from anywhere within the convex hull of this domain (Fig. 8). When this domain is highly non-convex, a large part of its convex hull lies outside of it. If $\langle \boldsymbol{m} \rangle$ does not correspond to any realizable graphs, the distribution must be degenerate. It is important to note though that degeneracy will also occur in situation when $\langle \boldsymbol{m} \rangle$ does correpond to a network, as shown in Fig. 5.

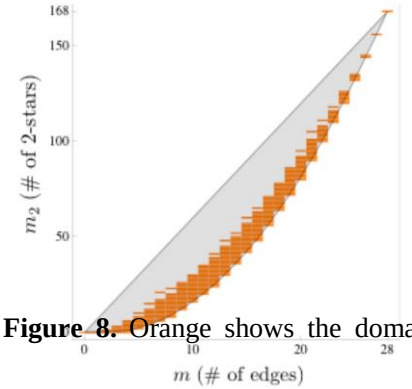


Figure 8. Orange shows the domain of realizable graphs for the measures $\boldsymbol{m}$ = no. of edges and $\boldsymbol{m_2}$ = no. of two-stars. The convex hull of this domain is shaded in grey. The average $(\langle \boldsymbol{m} \rangle, \langle \boldsymbol{m_2} \rangle)$ can take values from anywhere within the convex hull.

Avoiding degeneracy

We proposed a solution for avoiding degeneracy by replacing the constrained graph measures $\boldsymbol{m}$ with a one-to-one transformed alternative $\boldsymbol{\xi} = F(\boldsymbol{m})$ in such a way as to make the counting function $\mathcal{N}(\boldsymbol{\xi}) = \mathcal{N}(F^{-1}(\boldsymbol{\xi})) = \mathcal{N}(\boldsymbol{m})$ log-concave. In many practical situations this is possible by individually transforming each measure as $\xi_1 = F_1(m_1)$, $\xi_2 = F_2(m_2)$, …, $\xi_k = F_k(m_k)$.

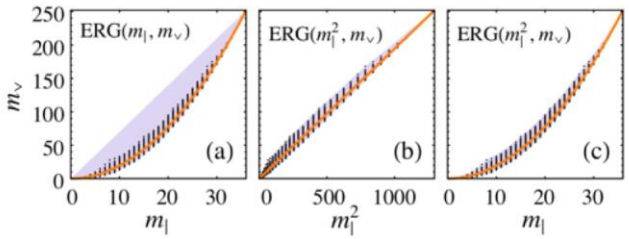


Figure 9. Black dots show the edge count and two-star count of graphs. The orangle line shows an approximate relationship between these two measures. The purpe shading shows the domain of the $(\langle m_1 \rangle, \langle m_\vee \rangle)$ averages.

(a) The ERG$!m_1, m_\vee!$ model is degenerate and averages will often not correspond to actual networks (b) and (c) In the ERG$!m_1^2, m_\vee!$ model most averages can be realized.

Let us illustrate this through the example of the ERG $(m_1, m_\vee)$ model. It can be shown that the approximate relationship $m_1 \propto m_\vee$ holds between the edge-count and two-star-count of graphs (orange line on Fig 9a). This causes the domain of $\mathcal{N}(m_1, m_\vee)$ to be crescent shaped, i.e. geometrically concave. Let us choose $\boldsymbol{\xi}$ so that the relationship between ξ_1 and $\xi_\vee$ is linear. After the transformation $\xi_1 = m_1$, $\xi_\vee = m_\vee$ we obtain a counting function $\bar{\mathcal{N}}(\xi_1, \xi_\vee)$ whose domain is approximately geometrically convex (Fig. 9b), which is a necessary condition for $\bar{\mathcal{N}}(\xi_1, \xi_\vee)$ to be log-concave. The ERG$(\xi_1, \xi_\vee)$ model does indeed turn out to produce non-degenerate distributions for most valid input constraints.

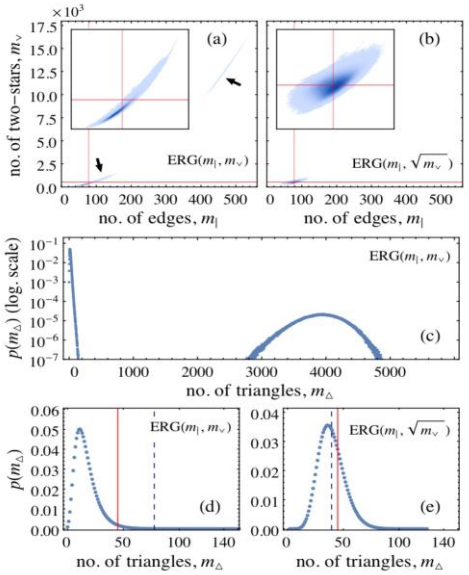


Figure 10. (a) Fitting **ERG**$!\boldsymbol{m}_1, \boldsymbol{m}_\vee!$ to Zachary's karate club dataset produces a degenerate distribuion. (b) Fitting **ERG**$!\boldsymbol{m}_1, \sqrt{\boldsymbol{m}_\vee}!$ avoids degeneracy. (c) and (d) **ERG**$!\boldsymbol{m}_1, \boldsymbol{m}_\vee!$ doesn't predict triangle counts well. (e) **ERG**$!\boldsymbol{m}_1, \sqrt{\boldsymbol{m}_\vee}!$ can predict triangle counts.

To apply this procedure for models constructed with two arbitrary graph measures, we must first map the relationship between the two measures. This may be done either analytically when possible, but can also be carried out using numerical methods. While this method does replace the graph measures $\boldsymbol{m}$ with an alternative set $\boldsymbol{\xi}$, there is a direct one-to-one relationship between these two. This means that if $\boldsymbol{m}$ has a simple intuitive interpretation, so does $\boldsymbol{\xi}$.

Numerical implementation

To fit an ERG model to some data $m^!$, we must numerically solve the equation $\langle m \rangle(\beta) = m^!$. Both of these are non-trivial tasks. carry out two tasks: (1) compute the average $\langle m \rangle(\beta)$ as a function of the parameter β; (2)

To compute the average, we implemented a fast Metropolis-Hastings sampler which can use the following graph measures: edge-count, two-star count and triangle count. To efficiently sample the space of graphs, it uses two basic MCMC moves: adding or removing a graph edge with equal probabilities; or taking the complement of the graph. These moves allow sampling even from degenerate distributions for as long as we are working with relatively small graphs. To invert the $\langle m \rangle \leftrightarrow \beta$ relationship we must use a numerical equation solver. Since $\langle m \rangle$ is computed with a MCMC sampler (i.e. a stochastic method), we cannot use standard numerical solvers and need to work with stochastic root finding methods instead. We used a combination of probabilistic bisection and a variant of the Robbins-Monro algorithm to create a customized stochastic solver.

Validation

We validated the method by fitting ERG models to two experimental datasets: Zachary's well know karate club social network data (34 nodes) [Zachary1977] and a collaboration network of jazz musicians (198 nodes) [Gleiser2003]. When fitting the ERG(m_1, m_∇) model to Zachary's karate club data, we obtain the parameters $\beta_1^* = 2.610$, $\beta_\nabla^* = -0.08125$ and a degenerate distribution with two well-separated peaks (Fig 10a). The predictions of this model for the triangle counts is inaccurate (Fig 10c and 10d). Exploiting the approximate relationship $m_\nabla \propto m_1^2$ between these measures, we can construct the model ERG$(m_1, m_\nabla^{1/2})$, which is expected not to be degenerate. Fitting this model to the empirical network yields the parameters $\beta_1 = 3.625$ and $\beta_\nabla = -7.998$ and a single-peaked, non-degenerate distribution (Fig 10b). This model predicts the triangle count well (Fig. 10e), giving an average triangle count of 39.5, while the actual triangle count of the network is 45. The value 39.5 has a high probability according to the triangle distribution obtained from this model (Fig 10e).

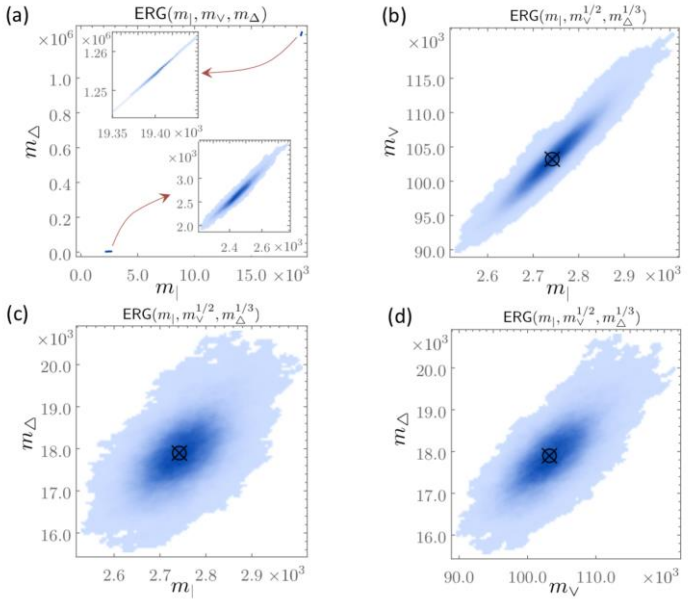


Figure 11. (a) Attempting to fit ERG$(m_1, m_\nabla, m_\Delta)$ to the jazz musicians network results in degeneracy. The remaining panels show the distributions (b) $p(m_1, m_\nabla)$, (c) $p(m_1, m_\Delta)$ and (d) $p(m_\nabla, m_\Delta)$ after fitting to ERG$(m_1, m_\nabla^{1/2}, m_\Delta^{1/3})$. The crosses show the locations of the average values.

Attempting to fit the model ERG($m_{-}, m_{\vee}, m_{\Delta}$) to the jazz musicians network fails because the distribution becomes degenerate (Fig 11a). For this larger network the averages $\langle m_{|} \rangle$, $\langle m_{\vee} \rangle$ and $\langle m_{\Delta} \rangle$ cannot be computed in a reasonable time using MCMC simulations when the distribution is degenerate. We constructed a model ERG($m_{|}, m_{\vee}^{!/!}, m^{\blacktriangle/!}$) where the relationship between the transformed graph measures is approximately linear. This model does not yield degenerate distributions when fitting it to the jazz musicians network. We obtain the fitting parameters

$\beta_{|} = 1.82$, $\beta_{\vee} = 12.5$ and $\beta_{\Delta} = -313$ and the distributions shown in Fig. 11bcd.

Applications of ERGs 1: Cortical Brain Networks

Brain neuronal networks provide a natural application for network modeling approaches and in particular for maximum entropy based models. The reason is that the data available from experiments is always noisy and most of the time incomplete. In such cases methods that extract information from the data and do not introduce any biases are necessary make reliable predictions about the system from which the data originates. So far, the only principled method that does that is Jaynes' Maximum entropy approach. Although his approach goes beyond networks, it has increasingly been receiving applications in the field of networks. Our goal was to develop and apply exponential random graphs for brain cortical networks, make predictions using those and then validate those predictions against experimental data. We have developed such model for macaque interareal cortical network datasets obtained from retrograde tracer experiments by a collaborating neuroanatomist team in Lyon, France, that was published in Science, see Ref. **[p37]**. We have also investigated whether the lessons learned from our work on macaque brains could be valid (and if so to what extent) for other mammals. However, this needed additional data. Only recently, two neuroscience experimental teams, one from University of South California (USC, Dong et. al) and another from the Allen Brain Institute (ABI) have generated similar cortical network datasets in the mouse cortex, using anterograde and retrograde tracing. It is important to mention that all existing papers of cortical networks have been of a descriptive nature, largely typical to biological papers, and there were no models of cortical networks with predictive power. Our work is state-of-the-art, we are the leading group in the world to present interareal (between functional areas) cortical network models with predictive capability. Our models are *not based* on fitting parameters, but on understanding of the essential biophysical constraints, and in particular the role of wiring costs and cortical geometry in generating cortical interareal connectivity networks. Our modeling methodology is based on the maximum entropy principle and in particular using exponential random graph models. In our studies on the macaque brain, the analysis of the weights and geometrical properties of the inter-areal connection pathways has revealed strong structural specificity; connection weights exhibit a heavy-tailed lognormal distribution spanning five orders of magnitude and conform to a distance rule reflecting exponential decline with the physical distance between areas. This exponential distance rule (EDR) has allowed us to define a single-parameter random graph model that predicts numerous features of the macaque cortical network: (i) Structural heterogeneity reflected in a dense network core; (ii) Global and local binary properties; (iii) Global and local weight-based communication efficiencies and (iv) Overall wire-length minimization.

When have applied our methodology on the mouse datasets from USC and ABI (which is a smooth, i.e., non-folded brain, unlike the macaque). We have found that just as in the macaque, there is an exponential cost to long-distance wiring, however, with a different decay rate. However, the corresponding EDR model also describes faithfully the network properties of the cortical data

network coming from the experiments, without any fitting parameters. Some of the findings are

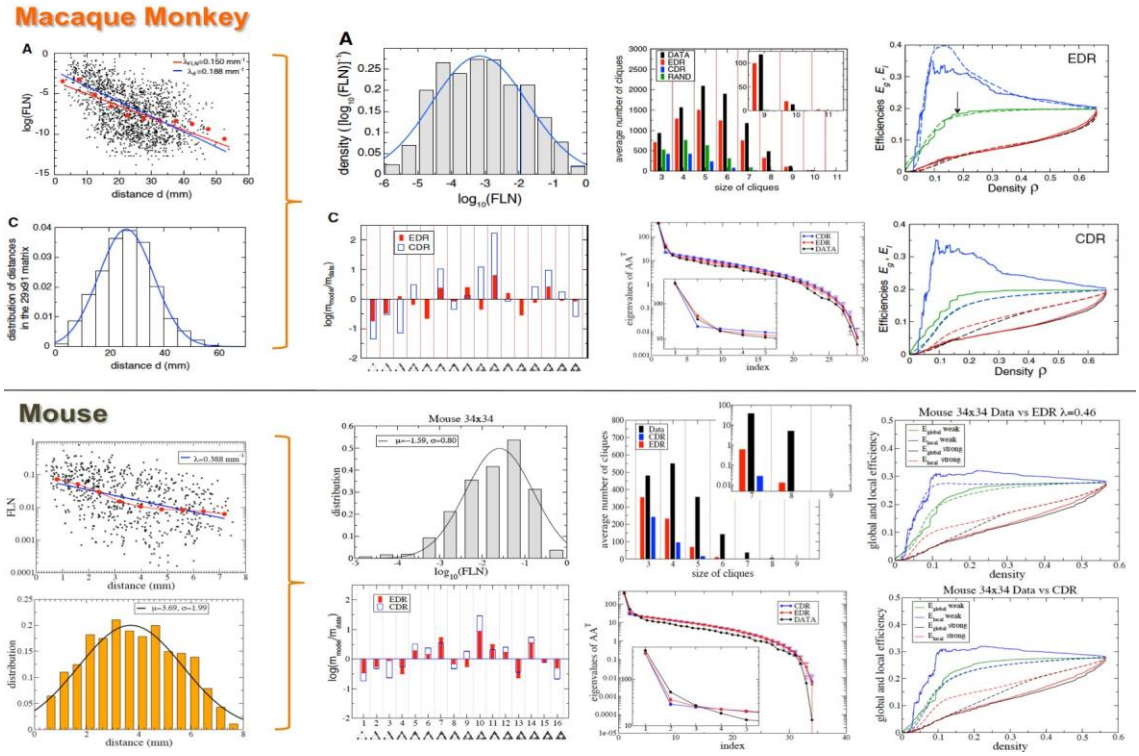


Figure 12. Comparison between various graph measures as obtained from the data network and the network models based on the exponential distance rule hypothesis (left-most column) in the macaque **[p37]** and mouse (recent work, [p).

exhibited in Fig. 12, comparing various binary and weighted graph theoretical measures between macaque and mouse brains. The success of the mammalian class as witnessed by its adaptation to diverse habitats and life styles is in part attributed to the behavioral flexibility ensured by the neocortex. The modulation of corticogenesis, lead to extant mammals exhibiting diverse morphologies and a large 5-order magnitude range of brain size, going from small-brained mammals that include miniaturization of ancestral forms to the expansion and additional arealization that characterizes primates. Our results show that the EDR and the distribution of distances are significant determinants of the inter-areal network of two representative species of this group.

Our network models use biophysical constraints (geometry and physics) as the fundamental mechanisms for graph generation, and essentially, they represent all other variables with uniform random variables in order to avoid introducing any biases. The degree to which the predictions from the models agree with experimental data is a measure in which the mechanisms and hypotheses introduced into the models are true determinants of the observed network structure. The fact that these EDR models indeed reproduce with good approximation the experimental data shows that these mechanisms play a fundamental role in the construction of brain cortical networks. In parallel with our cortical network simulations we have also performed analytic calculations that use morphological measures of the brain to determine the decay rate of the wiring cost with distance. These morphological measures include the mass of the white matter and the gray matter, the surface area of the gray-matter white-matter interface, the gyrification index and the smallest and largest

white matter projection distances. Our formulas capture the decay rate values to within the experimental error bars. These results are being submitted for publication.

Applications of ERGs 2: Predicting Flows in Roadway Networks

Our objective was, again, by using the maximum entropy principle approach to develop fitting parameter-free, principled models to predict dynamic features of networks, and in particular, network flows. Here we focused on the traffic flow prediction problem in spatial networks, and in particular in roadway networks and validated our results using US highway network and traffic data from MIT's arcgis database. Understanding flows in spatial networks driven by human mobility has many important consequences: it would enable us to connect throughput properties with demographic factors and network structure; it would inform urban planning; help forecast the spatio-temporal evolution of epidemic patterns, help assess network vulnerabilities, and allow the prediction of changes in the wake of catastrophic events, such as weapons of mass destruction effects, where a significant geolocalized section of the network becomes inaccessible and inoperable.

This work was motivated by the need to understand, model and predict flows in complex networks and in particular human mobility in spatial networks. The ability to predict network flows in such systems is far behind our ability to determine the flows in, for e.g., electric and electronic circuits, essentially due to the social component driving the transportation dynamics. A number of recent publications (several in *Nature* journals), however, have shown that aggregated measures of human mobility follow statistically reproducible behaviors, indicating that first-principles based description and understanding is possible for flows in large and complex infrastructure networks, as we indeed demonstrated it also in our work.

The quantitative prediction of commuter flows through the network requires the solution of two fundamental problems. One is of mainly socio-demographic nature and it entails determining the number of individuals travelling between two given locations (The Mobility Law) and the other is a network distribution problem (The Flux Distribution problem) and it entails assigning the network paths to the individuals that are travelling between the locations. We have solved both problems and provided a model for traffic flows that is predictive and not based on fitting parameters. The results have been published in the journal *Nature Communications*, **[p27]**.

Since this model is based on first principles, it can be used to predict commuter flows on any roadway network. In particular it can be used to predict flows when the existing roadway network suffers a change due to catastrophic events (earthquakes, land slides, etc.) or weapons of mass destruction attacks. In this case a geo-localized portion of the network becomes unusable for further transport, and this leads to reorganization of the flows. A key question that arises is: How far the effects are felt in the network flows from the location of the damage? Is there a Newton's cradle effect that is when long-range effects appear due to localized damages? Since our model is principled, it can then be rerun with modified network and population structure and the changes around the damage assessed. And indeed, in spite the fact that in roadway networks, as it is the case in spatial networks, there are no shortcuts (like in airline networks), yet significant changes can be

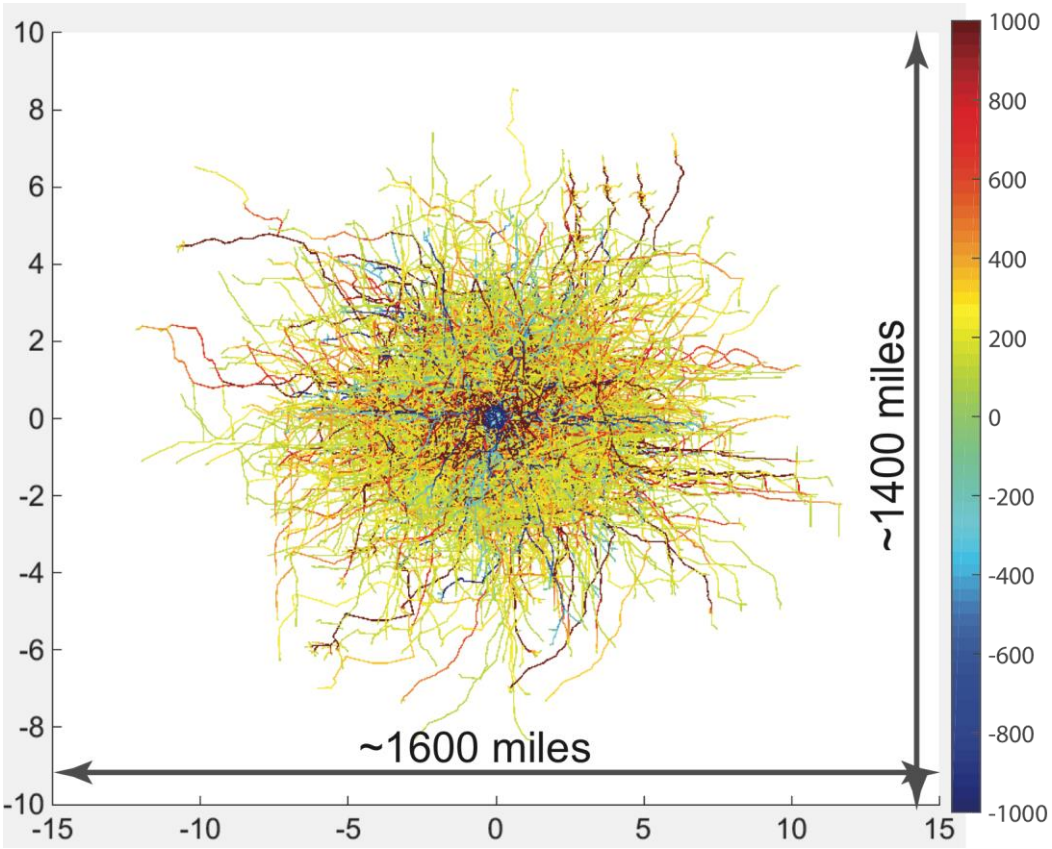


Figure 13. Distribution of traffic changes in the wake of a geolocalized damage that removes the network and the population within a 30km radius area. This was done for 556 damage centers chosen randomly from the continental US. Network fragility hotspots cause traffic pattern changes that are long-range.

felt long distances from the damage center, for certain damage centers (fragility hotspots). Identifying these fragility hotspots is a critical step in ensuring the resilience of the transportation network to external attacks and geo-localized damages. Fig. 13 shows the distribution of traffic changes (color map) around the damage center for 556 randomly chosen damage centers. The damage centers have been overlaid onto the same point. The damages were induced by removing a 30km radius circular region from the network around the damage center. One can see that the effects can be felt out to thousands of kms, almost half the size of the whole US. The reason for the existence of such long-range effects is that transport is non-local and that the same road segment can be used by a large number of source-destination pairs. Once road segments are removed, the lowest

cost paths are redistributed, and this can cause a long-range change in the traffic distribution for certain locations in the network (fragility hotspots). The mathematical description of these long-range propagating effects are under development, and will be submitted soon for publication.

Laplacian Spectra of Ensembles of Random Geometric Graphs

Using both analytic and computational methods, we have studied the Laplacian spectra of random geometric graphs (RGGs) with a goal of relating those spectral properties to the structure of the graphs. RGGs are simple models of spatial networks, i.e. networks in which the nodes have a topological location and a distance between nodes, independent of the links of the network, can be measured. Spatial networks arise in transportation and power grid networks, as well as cell phone and ad-hoc sensor networks. The RGGs we consider are created by randomly placing nodes in some topological space, with some boundary conditions, and then connecting all nodes that lie within some fixed distance from each other. The spectra of the graph Laplacian are the eigenvalues of the discrete generalization of the Laplacian operator on the graph. The eigenvalue correspond to modes of diffusion and also of the propagation of waves on the network. They also describe certain structural properties including the robustness of the connectivity. We consider the ensemble-averaged spectra.

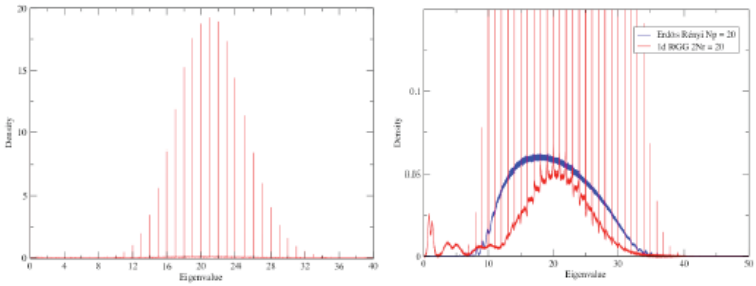


Figure 14. Eigenvalue Spectrum of an Ensemble of RGGs. Results are shown for one-dimensional RGGs on a circle. The figure on the left shows the discrete part of the spectrum, which consists of integer-valued peaks. The figure on the right, which is an enlargement of the left figure, shows the continuous part of the spectrum. The blue curve shows the corresponding spectrum for Erdos-Renyi graphs.

A striking contrast between the spectra of RGGs and of non-spatial random networks, such as Erdos-Renyi networks, is that there are two parts that persist in the large network limit: a discrete part and a continuous part. See Fig. 14. The discrete part, which consists of integer-valued eigenvalues, vanishes in the large network limit (with fixed connection radius r) of non-spatial nets, but we have analytically shown that they make up a finite fraction of the total number of eigenvalues in spatial networks. We have also shown that these integer eigenvalues are due to mesoscopic structures within the network that correspond to graph automorphisms associated with permutation symmetries of the graphs. Specifically, we identify two types and

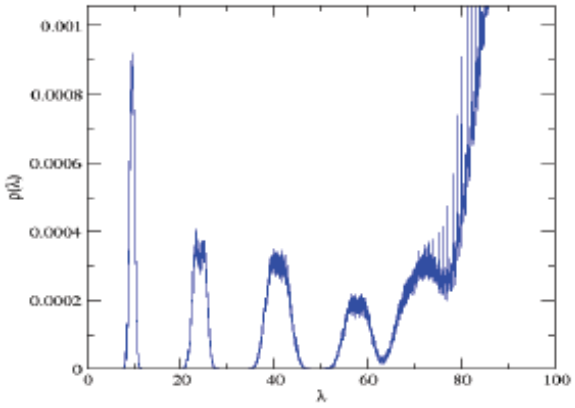


Figure 15. Low end of the eigenvalue spectrum of Laplacian matrices of RGGs on two-dimensional disks with open boundary conditions. Note the series of eigenvalues that separate from the bulk of the distribution, the left tail of which is shown in the right side of the figure.

calculate their expected multiplicity. These two types correspond to the simplest possible symmetric meso-scale structures. The eigenvectors of the corresponding modes are localized, and so in principle they can be excited without exciting the other modes. We have analytically calculated the expected distribution of these integer eigenvalues for RGGs in a variety of different topologies of embedding spaces and confirmed our analytic findings with computer simulations.

The continuous part of the spectrum of RGGs also has interesting properties. Most notably, as the model parameters are varied, finite sets of eigenvalues will separate from the main, bulk continuous distribution, as can be seen in Fig. 15 and in the right figure of Fig. 14. The existence of these separated eigenvalues leads to localization phenomena related to Anderson localization in electronic states of crystals. By approximating RGGs with random continuous media, we have been able to calculate the number of eigenvalues that separate off in different topologies of embedding spaces. Furthermore, by approximating RGGs with lattices, we have been able to determine how the leading, separated eigenvalues scale with the model parameters. Understanding this is particularly important, especially for μ_{N-1} the smallest nonzero eigenvalue that is known as the spectral radius. In particular, we find that

$$\mu_{N-1} = k + 1 - (k+1)^{1-1/d} \frac{\sin\left[(k+1)^{1/d} \frac{\pi}{N^{1/d}}\right]}{\sin \frac{\pi}{N^{1/d}}} \approx \frac{1}{6} \left(\frac{\pi}{N^{1/d}}\right)^2 (k+1)^{1+2/d}$$

where N is the number of nodes in the network and k is the average degree of the nodes. This eigenvalue provides bounds for both the node and link connectivity of the network. This result has application for predicting the robustness of ad-hoc sensor networks to node failure or attack.

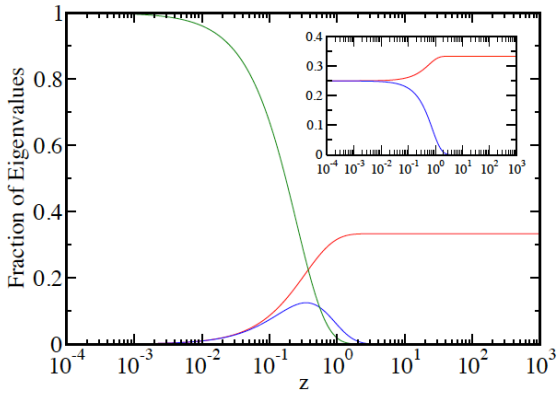


Figure 16. Fraction of eigenvalues due to Type-I orbits (red), Type-II orbits (blue), and in the zero eigenvalue condensate (green) in the extensive large network limit as a function of z = Nr. Inset shows the fraction of eigenvalues not in the condensate that are due to Type-I (red) and Type-II (blue) orbits.

Perhaps a more important thermodynamic limit though is the extensive limit in which N → ∞ while the average degree Nr = z is constant. In this case, for small z, we find a phenomenon reminiscent of Bose-Einstein condensation in the accumulation of zero eigenvalues. Fig. 16 shows that, in this limit, for z ≫ 1, as in the nonextensive limit with fixed r, a third of the eigenvalues are due to Type-I orbits, while virtually none are due to Type-II orbits and the zero eigenvalue condensate is empty. However, near the giant component transition, z **1** 1, the situation changes. Here the fraction of eigenvalues due to Type-I starts to decrease, the condensate starts to fill, and the fraction of eigenvalues due to Type-II orbits reaches a maximum. For z ≪ 1, the condensate absorbs almost all of the eigenvalues, while the fraction of the eigenvalues that are due to simple orbits

vanishes. However, as shown in the inset of Fig. 16, even for $z \ll 1$, a substantial fraction of the eigenvalues that are not in the condensate are always eigenvalues due to simple orbits. In this limit Type-I and Type-II orbits each produce a quarter of the eigenvalues that are not in the condensate.

The existence of such a large number of orbits in spatial networks can also be used to simplify the analysis of their behavior. By considering a quotient graph the integer eigenvalues can be removed, leaving only the continuous part of the spectrum. The continuous part of the spectrum describes much of the important properties of the network's behavior. For example, as mentioned earlier, the smallest nonzero eigenvalue in graph Laplacian matrices determines the number of vertices or edges that must be cut to sever the network. Also, its eigenvector can be used to partition the network into communities. Notice from Fig. 14 that, in the case shown, pairs of eigenvalues split-off or separate from the bulk continuous distribution at the small end of the spectrum. These separated eigenvalues include the smallest nonzero one. Ongoing work indicates that the number of eigenvalues that split off together from the bulk continuous distribution can be deduced by approximating the graph Laplacian with a continuous Laplacian operator corresponding to disordered random media and considering the degeneracy of the modes with smallest eigenvalues. The results on the multiplicity of graph orbits in RGGs is described in **[p21]**.

An exotic phase transition in a minimal model of a social network

In common descriptions of phase transitions, first order transitions are characterized by discontinuous jumps in the order parameter and normal fluctuations, while second order transitions are associated with no jumps and anomalous fluctuations. Outside this paradigm are systems exhibiting 'mixed order transitions', displaying a mixture of these characteristics. When the jump is maximal and the fluctuations range over the entire range of allowed values, the behavior has been coined as an 'extreme Thouless effect'.

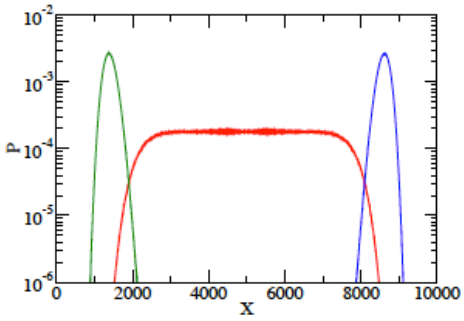


Figure 17. Probability distribution of the number of links X for three cases with (N_I, N_E) near/at criticality, shown in green (101, 99), red (100, 100), and blue (99, 101). Note the extreme change that happens when only one node switches temperament from introverted to extroverted, or vice-versa.

As reported in the paper **[p19]**, we have found evidence of such a phenomenon in the context of dynamic, social networks. Defined by minimal rules of evolution, the model describes a population of extreme introverts (I) and extroverts (E), who prefer to have contacts with, respectively, no one or everyone. From the dynamics, we have derived an exact distribution of microstates in the stationary state. With only two control parameters, $N_{I,E}$ (the number of each subgroup), we have studied collective variables of interest, e.g., X, the total number of I-E links and the degree distributions. Using simulations and mean-field theory, we found evidence that this system displays an extreme Thouless effect. Specifically, as shown in Fig. 17, the fraction $X/(N_I N_E)$ jumps from 0 to 1 (in the thermodynamic limit) when N_I crosses N_E, while all values appear with equal probability at $N_I = N_E$.

T2. Finding Structures of Interest

Optimizing community detection

Identifying structure within a network and in particular in large-scale real-world network data is one of the central topics in Network Science. Many networks have been found to possess a modular structure that can influence the dynamical processes supported by the network, affecting for example synchronization behavior, percolation properties, and the spreading of epidemics. A commonly used indicator of the prominence of modular, or community, structure in a complex network is its modularity *Q*. For a given partitioning of the vertices, modularity is defined as the fraction of edges that occur within partitions minus the fraction that would occur on average in those same partitions within an ensemble of random networks. The modularity of a network is the maximum modularity of any vertex partition. This is an intuitively appealing way of identifying community structure within a network, and has become a standard way of doing so. However, characterizing community structure within networks using modularity presents challenges and issues.

The main challenge is that finding the vertex partition that maximizes modularity is an NP-hard computational problem. For practical applications, therefore, it is important to have a fast algorithm that will complete in polynomial time and which will produce an accurate estimate of the modularity of any given network. Among the issues associated with using modularity to find community structure in networks is that it can be difficult to interpret the results. In particular, in general, modularity itself does not allow for the quantitative comparison of the modular structure in different networks. Among networks with the same number of vertices and edges, a higher modularity does indicate a more modular network structure, but this is not necessarily the case when networks with different number of vertices and/or edges are compared.

We have addressed both the challenge of developing a fast and accurate algorithm for finding the network partition that maximizes modularity and the issue that modularity itself is not a useful measure for making direct quantitative comparison of the modular structure in networks that have different number of vertices and/or edges. Our results provide practical solutions to both the main challenge concerning detecting community structure in complex networks and the important issue of comparing the community structure in different networks.

There has been considerable interest in finding a fast and accurate algorithm for finding the partition of the vertex set of any given network that maximizes modularity, and thereby determining the modularity of the network. This is a difficult, NP-hard computational problem. There have been numerous algorithms developed for this purpose. Perhaps the best known is the, so-called, leading eigenvalue method developed by Newman. This algorithm works by partitioning the vertex set through recursive bisectioning until no further improvement in modularity is found. In order to perform each bisection, the algorithm makes a guess at the best split of the vertices by finding the eigenvector of the largest, or leading, eigenvalue of the modularity matrix. This guess can be substantially improved by using a Kernighan-Lin type refinement step that was also introduced by Newman. A number of other modularity maximizing algorithms also work by recursively bisecting the vertex set.

However, in previous work [Sun2009] we showed that recursive bisectioning introduces constraints on the partitions considered, thereby limiting the accuracy of the algorithm. To solve this problem, we proposed an algorithm that included a different Kernighan-Lin type refinement step that removed the constraints imposed through recursive bisectioning. At the time it was introduced, our algorithm outperformed all other known fast modularity maximizing algorithms. In the mean time though, a couple of new algorithms have been introduced that outperform our earlier algorithm.

Furthermore, until now there has been no way to use modularity to make direct quantitative comparison of the modular structure in networks that have different number of vertices and/or edges. The only way such comparisons can currently be done is to use other network measures that are less

intuitively appealing than modularity.

As detailed in our paper **[p23]**, we have developed a fast spectral algorithm that improves on our previous modularity maximizing algorithm. It is based on the leading eigenvalue method with a Kernighan-Lin type refinement introduced by Newman, but improves on that algorithm by using a combination of an agglomeration step and our previously developed different Kernighan-Lin type refinement. This algorithm improves on existing ones by removing constraints on the partitions considered, similar to the way that our earlier algorithm obtained improved results. Our earlier algorithm had difficulty finding the modularity maximizing vertex partition in larger networks. By including an agglomeration step that can combine two partitions, we now can achieve highly accurate results for networks with even tens of thousands of vertices. As discussed in below, our algorithm outperforms all other known fast modularity maximizing algorithms when applied to the find the modularity of a set of commonly used test networks.

In order to allow the quantitative comparison of the modularity in different networks, we have developed a practical way of measuring the "effect size" of the modularity of a network. Using Erdos-Renyi networks from the $G(N,p)$ ensemble that have the same number of vertices and on average the same number of edges as the network being considered as a null model, we have found an analytic expression for the modularity z-score, which quantifies modularity effect size. The z-score is defined as the number of standard deviations away from null model's expectations for the modularity that the modularity of a network actually is. A positive z-score indicates a more modular community than would be expected in a comparable random network, while a negative z-score indicates a less modular community structure than expected in a comparable random network. Our analytic expression for the modularity z-score is a function of a network's modularity and of the number of vertices and edges it has.

In order to obtain our analytic expression for modularity z-score, we performed an extensive numerical study of the distribution of modularity in ensembles of Erdos-Rényi networks. The accuracy of our new modularity maximizing algorithm made this study possible. Then, using our numerical results, we fitted finite-size corrections to theoretical predictions previously derived for the mean of the distribution and to a novel expression we derived for the variance of the distribution, both of which are valid in the large, dense network limit. The finite-size corrected analytic expression we thus obtain can be used to simply map a modularity value into a z-score measure of the effect size of modularity. As we demonstrate in our preprint, our expression is accurate for networks of any size that have an average degree of 1 or more. The modularity z-score can be used to quantitatively compare the modularity of different networks, including those with different numbers of vertices and/or edges.

In order to validate the accuracy of our modularity maximizing algorithm, we used it to find the modularity of a number of real-world commonly studied benchmark networks. The results are shown in Table 1.

Network	Vertices	Edges	Q	z-score	Q_{pub}
Karate Club	34	78	0.4198	1.68	0.4198
Dolphins	62	159	0.5285	5.76	0.5276
Political Books	105	441	0.5272	18.27	0.5272
David Copperfield	112	425	0.3134	-3.51	0.3051
Jazz	198	2742	0.4454	108.91	0.4454
C. Elegans	453	2025	0.4526	21.97	0.4522
Tarragona emails	1133	5045	0.5827	70.89	0.5825
Key Signing	10680	24316	0.8837	-144.17	0.884

Table 1. Algorithm validation. The comparison between the maximum modularity found by our algorithm (Q) and the best published result (Q_{pub}) shows that no other modularity maximizing scheme performs better than our method. The benchmark networks used are, in order, the social network in an American karate gym, the social network of a community of dolphins in New Zealand, the network of co-purchases of political books on Amazon.com in 2004, the word adjacency network in David Copperfield, a collaboration network between jazz musicians, the metabolic network in C. Elegans, the network of emails exchanged between members of the Universitat Rovira i Virgili in Tarragona, and a network of trust in cryptographic key signing.

A description of the networks (data sets) used for our validation is given in the caption. As the Table shows, we find that no other currently known fast modularity maximizing algorithm performs better than our new algorithm on any of the networks studied. Note that the maximum possible value of modularity is 1.0. Thus, the networks considered vary from ones with relatively low modularity (the word adjacency network of in David Copperfield) to ones with quite high modularity (a trust network in cryptographic key signing). The networks also range from ones with tens of nodes to ten thousand nodes.

Also shown in Table 1 are the modularity *z*-scores of the real-world test networks we used to validate our algorithm. The conversion from modularity value to the *z*-score of modularity effect size we have established is particularly noteworthy. Because of it, for the first time, one can easily make direct quantitative comparison of the modularity in networks with different numbers of vertices and/or edges. This allows a new form of comparative network analysis. For example, note that most of the networks have a *z*-score much greater than 1 and, thus, a substantial modular structure, with the collaboration network of Jazz musicians being by far the most modular of those studied. However, the network of trust in cryptographic key signing has a large negative *z*-score and thus is substantially less modular than a comparable ER network. This indicates that the links in that network are much more evenly distributed than expected if it were random. Furthermore, the word adjacency network in David Copperfield has a slightly less modular structure than expected if random, and the Karate Club network has only a very weak modular structure. This form of analysis is clearly much more informative than one that considers modularity alone. The difference is particularly striking when one considers the network of trust in cryptographic key signing. It has a very large modularity value, but this value does not indicate that it has a highly modular structure relative to a random network. In fact, it has much *less* community structure than comparable random networks. In on-going work we seek to extend our algorithm for modularity maximization to weighted and bipartite networks, and to develop modularity z-score measures applicable to those types of networks.

Lovasz Local Lemma and its applications in network science

The Lovasz Local Lemma (LLL) is a powerful probabilistic tool in combinatorics, used for existence proofs, which is said to find the proverbial needle in a haystack. LLL sets a lower bound for the probability that none of certain events happen. The use of LLL requires a dependency graph defined on the set of events involved, expressing lots of mutual independence. The LLL extends to a more general setting when a ‘‘negative dependency graph’’ is present [Alon2000]. This extended version has not had many applications as it difficult to identify negative dependency graphs that are not dependency graphs. Earlier Lu and Szekely [Lu2007] defined proper negative dependency graphs in the space of random injections between two sets, in an alternative description in the space of perfect matchings in a complete bipartite graph.The reported paper **[p14]**, last revised in 2014, extended this to the space of perfect matchings in a complete graph. In these two settings **[p14]** also

provided close upper bounds for the probability that none of the events happen. Note that partial matching are exactly the objects that the configuration model of network science considers. In this way **[p14]** reproved or proved asymptotic enumeration results for regular graphs, regular graphs with excluded cycle conditions, and graphs with prescribed degree sequences and girth conditions. This is relevant for the goal Constructing, sampling from graph ensembles obeying given constraints as counting and sampling are often closely related problems.

[p43] defined more negative dependency graphs, based on matchings in hypergraphs instead of graphs, and on forests. This opened up the way for using LLL for hypergraph enumeration as well. This was done in [Mohr13] for 3-uniform hypergraphs, and is under work in general **[p50]**. Threshold functions for balanced subgraphs in the Erdos-Renyi model are presented in every text for random graphs. The reported paper **[p45]** extended asymptotic formulas in **[p14]** from excluded cycles to some excluded balanced subgraphs (note that cycles are balanced!), and derived asymptotic results on the probability that a random regular multigraph from the configuration model contains at least one from a fixed family of balanced subgraphs.

Scaling of Minimum Dominating Sets in Scale-Free Network Ensembles

A central issue arising in the context of networked systems is the ability to efficiently control, monitor, detect, or influence the behavior of the constituent nodes of a network. In static networks or slowly evolving networks, a solution to this problem often involves computing a dominating set of the network. A dominating set of a network (graph) G with node set V is a subset of nodes $S \subseteq V$ such that every node not in S is adjacent to at least one node in S. Example problems in whose solution the dominating set (or some variant of it) has been shown to play a part include optimal sensor placement for disease outbreak detection, controllability of networks and social influence

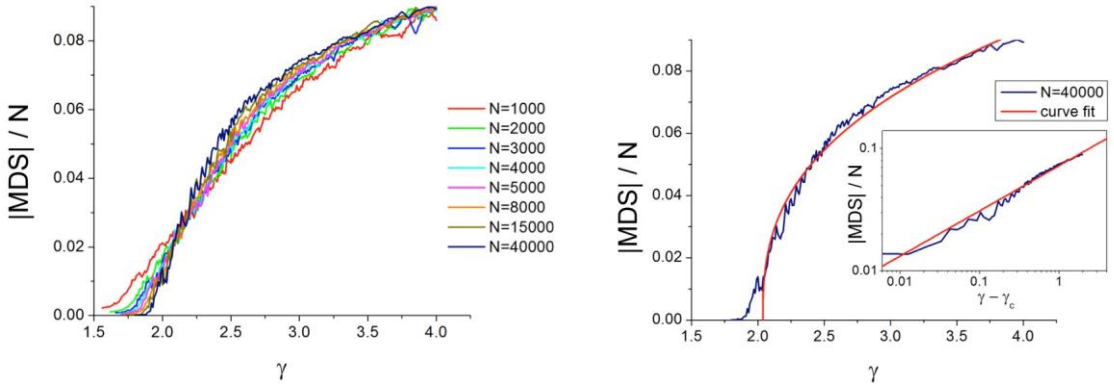


Figure 18. Domination transition in scale-free HHMC network ensembles with no structural cut-offs (k_{max} = N–1). **(a)** shows the scaled MDS size vs. γ with $\langle k \rangle$ =14 for various system sizes**. (b)** Scaled MDS size for the largest network and the best-fit power-law (solid red curve) in the vicinity of (and above) the transition point, $|MDS|/N \propto (\gamma - \gamma_c)^{\beta}$ with $\beta \approx 0.37$. Inset: same data as in (b) after shifting the horizontal axis and on log-log scales.

propagation. The smallest dominating set of G constitutes its minimum dominating set (MDS). Thus, the MDS of a network is the smallest subset of nodes such that every node of the network either belongs to this subset or is adjacent to at least one node in this set.

We studied and obtained the scaling behavior of the size of minimum dominating set (MDS) in scale-free network ensembles, with respect to network size N and power-law exponent γ, while keeping the average degree fixed. We considered ensembles generated by three different network construction methods, and we developed an efficient greedy algorithm to approximate the MDS

[p36]. With a structural cutoff imposed on the maximal degree ($k_{max}=\sqrt{N}$) we found linear scaling of the MDS size with respect to N in all the three network classes. Without any cutoff ($k_{max}=N$-1) two of the network classes display a transition at $\gamma\approx1.9$, with linear scaling above, and vanishingly weak dependence below (Fig. 18), but in the third network class we find linear scaling irrespective of γ. We found that the partial MDS, which dominates a given z<1 fraction of nodes, displays essentially the same scaling behavior as the MDS **[p36]**.

Theoretically, finding the exact MDS is an NP-hard problem, but it can be approximated reasonably well using fast algorithms. In order to utilize the MDS in real-world applications it is of paramount importance to be able to find an MDS efficiently. Our implementation of a greedy MDS search algorithm is designed with this priority in mind. We use bucket sort with hashed lists as buckets to sort and pick nodes for the MDS in $O(1)$ time, which allows us to find MDS approximations in $O(M)$ time (linear in the number of edges M) consistently for any network. This is far superior in applicability to large-scale networks, in contrast to the binary integer programming approach [Nacher2012], that takes an unknown number of $O(N^!)$ iterations, but inherently exponential in time, and if terminated early, provides no significantly better MDS than the greedy algorithm. Our results on the scaling of dominating sets have been published in **[p36]**.

Dominating Networks Using Generalized Probabilistic Methods

Probabilistic and cut-off dominating sets

Dominating sets play a critical role in the function and control of biological, social, and infrastructure networks. In particular, they provide key solutions to generalized network controllability, social influencing, optimal sensor placement for disease outbreak detection, observability of the power-grid, or finding high-impact optimized subsets in protein interaction networks. A dominating set is a set of nodes of a network (dominator nodes) such that each node

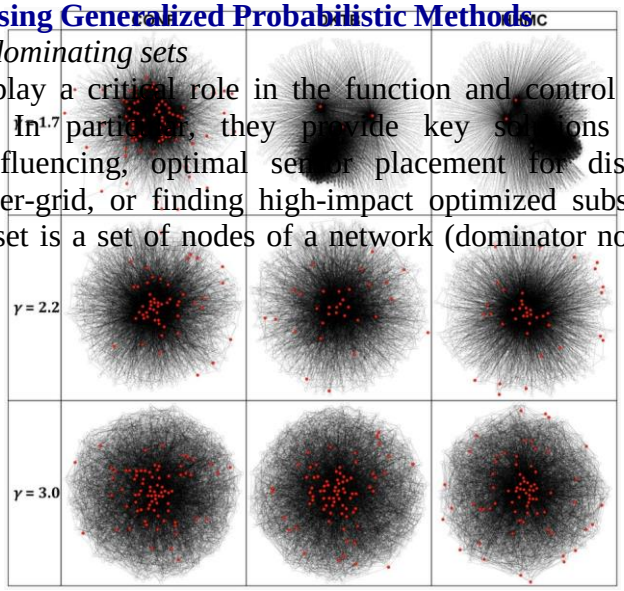


Figure 19. Visualization of typical scale-free networks of each type with no structural cut-offs (k_{max} = N–1) at three different power-law exponent values. In all networks, N=1,000 and $\langle k\rangle$ =14; the red-colored nodes belong to the MDS.

that is not in the set is adjacent to at least one dominator node. The smallest subset of nodes that can dominate the entire network is the minimum dominating set (MDS). Since node inclusions in the dominating set come at a certain cost, it is important to find the smallest MDS of a network for cost-efficient purposes. Finding the MDS is an NP-hard problem, and current research is focused on finding better approximations to the MDS. However, the sophisticated MDS search algorithms require rigorous knowledge of network structure and connection schemes, and obtaining such information on large-scale networks comes at high costs. Also, these algorithms have high computational time complexity. Thus it is of significant importance to develop strategies for finding optimal dominating sets that satisfy cost-efficiency demands in terms of dominating set size, computational time complexity and required network structure information.

Based on the classic graph theory approach for finding the upper bounds [Alon2000], we developed two novel probabilistic dominating set selection methods, applicable to heterogeneous networks, aiming to approximate the MDS size. We investigated the efficiency of these methods in comparison with the sequential greedy search algorithm that produces the MDS, and the deterministic degree-ranked dominating set (DDS) selection method, which sorts the nodes by their degree in non-increasing order, and adds the nodes, in this order, into the dominating set until the entire network is dominated. Classic probabilistic methods for finding dominating sets do not consider the heterogeneous nature of scale-free networks. Thus, our goal is to develop probabilistic methods that are applicable to heterogeneous networks. Therefore, in our first method we improve the classical probabilistic selection strategy by introducing in the selection probability a node degree-dependent criteria $p_i=(k_i/k_{max})^{\beta}$, where β is a parameter.

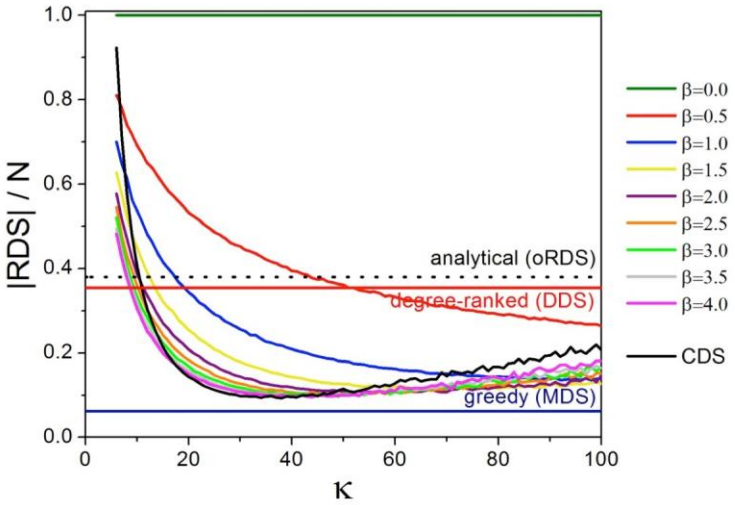


Figure 20. Cutoff dominating set as a function of κ degree cutoff parameter in the degree-dependent node selection probability. For comparison, curves of RDS are plotted for various β values. CDS corresponds to $\beta = \infty$.

Our numerical results revealed that the degree-dependent random dominating set (RDS) obtained in such manner results in a smaller dominating set size than the pure probabilistic one, and also outperforms the deterministic degree-ranked selection. Our second strategy is a limiting case of the RDS, where $\beta\to\infty$. This method is based on a degree cut-off; nodes having higher number of connections than this cutoff, are added into the cutoff dominating set (CDS). Our numerical results demonstrate that this method provides a significantly smaller dominating set than the classic

probabilistic method, RDS, and also outperforms the deterministic degree-ranked selection (Fig. 20). As a significant result, we found that the size of the CDS can closely approach that of the MDS provided by the sequential greedy algorithm, irrespective of particular topological properties of the scale-free network. In addition, the algorithm uses only local connectivity information, resulting in a highly cost-efficient method, and is also suitable for large network analysis.

In our ensemble-based numerical study we have employed the configuration model to generate scale-free networks from prescribed degree sequences. In addition, we generated two scale-free network subclasses: networks with no cutoff ($k_{max} = N$-1) and with imposed cutoff ($k_{max} = \sqrt{N}$), resulting in uncorrelated networks. We have investigated the behavior of four distinct dominating sets with respect to multiple network features (size, average degree, maximum degree cutoff, power-law exponent). We also tested our findings on several real-world social network samples, and found that they are in agreement with the results obtained on artificial networks. The real-world network samples were constructed using various collaboration datasets downloaded from the publically and freely available Stanford Dataset Collection. We hve validated the efficiency of our proposed dominating set selection strategies on various synthetic scale-free network samples and on several real-world social networks. Our next objective is to develop analytical estimates for our RDS and CDS methods to further validate our numerical results. In addition, finding analytical estimates would allow one to solve the estimate and find an accurate expected size of these dominating sets. Thus the analytical estimates would provide a powerful tool for finding the dominating set of a network without numerical calculations, and requiring only basic network property information.

We developed two novel probabilistic dominating set selection strategies and found that one of them produces the smallest dominating set size among probabilistic methods. In addition we have demonstrated that its size approximates the MDS, and this method outperforms the deterministic degree-ranked selection. Our developed CDS method satisfies the cost-efficiency demand, as it requires only local information. Using extensive numerical analysis on artificial scale-free network ensembles and various real-world social networks, we validated our conjectures on both synthetic and real networks.

Impact of network assortativity on network domination

Recent studies analyzed the scaling behavior of MDS in scale-free networks with a wide range of network sizes and degree exponents. It was found that the MDS size decreases as γ is lowered, and in certain special cases when the network structure allows the presence of $O(N)$ degree hubs (when $\gamma < 2$), the MDS size shows a transition from linear to $O(1)$ scaling with respect to network size, making these heterogeneous networks very easy to control. However, the impact of network assortativity, which is a fundamental property in real networks, has not been studied.

Therefore, we have investigated the impact of assortativity on the efficiency and size various approximate dominating sets in complex networks. In complex networked systems, mixing patterns are usually described by assortativity measures. A network is considered assortative if its nodes tend to connect to other nodes which have similar number of connections, while in a disassortative

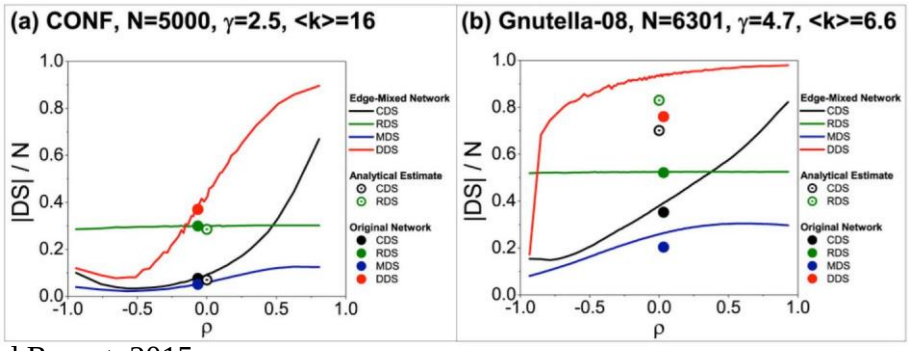


Figure 21. Networks with assortativity values different from the original network are obtained by guided edge-mixing with 2- swaps.

network the high degree nodes are adjacent to low degree nodes. Investigating the behavior of dominating sets with respect to assortativity is essential for deeper understanding of the network domination problem. Several studies conducted on real-world networks have shown that social systems are assortative, while technological ones exhibit disassortative behavior. Social psychology studies have shown that humans are more likely to establish a connection with individuals from the same social class, or with whom they share common interests, such as education or workplace. This tendency, named homophily, also governs the attachment rules in real-life social systems, and it is reflected in the mixing patterns of these networks, which are of significant importance in dynamical processes on social networks. Specific connectivity schemes affect influence propagation and epidemic spread, and is also responsible for Web page ranking and internet protocol performance.

We have developed and employed a new method to efficiently control assortativity in network ensembles. Using this technique, our goal is to provide a large-scale analysis on the behavior of various dominating sets, with respect to a wide range of network parameters, including assortativity. Finally, we also compared our findings on model scale-free networks and real-world network samples.

Using our edge-mixing method to control the assortativity of a network, we have compared the sizes of dominating sets as a function of assortativity, measured by Spearman's ρ. Fig. 21 shows our results for a synthetic network and a real social network.

As expected, the size of most dominating sets increase with higher assortativity, except for RDS with degree-independent selection probability. The most dramatic size increase is observed in DDS, which indicates that this method can only be considered viable in real-world applications for highly disassortative networks. Also, as the assortativity increases, CDS becomes larger than the simple RDS at a certain point, indicating that favoring high-degree nodes as dominators is not an effective strategy when the network is highly assortative. While the MDS size obtained by greedy search also increases with increasing assortativity, it shows the smallest increase, thus the advantage of greedy search over other methods is more pronounced.

Our numerical study of dominating set sizes with respect to assortativity revealed a general tendency that the dominating set becomes larger as assortativity increases. We can understand this easily. In case of a disassortative network, high degree nodes connect mostly to low degree nodes, therefore we can expect small dominating sets, due to efficient domination via high-degree nodes. In fact, when $\gamma < 2$ scale-free networks may become so disassortative that star subgraphs form and the size of MDS becomes $O(1)$. On the other hand, hubs are less effective in dominating assortative networks, since most of their connections are used to connect to other high degree nodes. Therefore, the impact of assortativity on each dominating set selection method depends on how much the method relies on high-degree nodes as dominators. This is why the degree-ranked selection shows the worst performance on highly assortative networks, followed by the degree-dependent RDS (and its limiting case, the CDS), which also favors high-degree nodes. Since technological scale-free networks tend to be disassortative, and although social networks tend to be assortative, extreme assortativity is rare, we can safely conclude that CDS is a viable alternative of greedy selection for most scale-free networks. Our work on these findings and on the general scaling properties of probabilistic dominating sets, in collaboration with É. Czabarka and L. Székely from USC, have been recently published in **[p30]**.

Stability of Dominating Sets in Complex Networks against Failures and Attacks

Attacks on complex networks and their structural effects have been extensively studied in the past, however no study has focused on the analysis of vulnerability of dominating sets against

various forms of network damage. Dominating sets play a critical role in network observability and controlling, therefore it is important to understand how robust these dominating sets are in case of random failure or targeted attack. Our aim is to study the resilience of various dominating sets against different network damage scenarios, and to develop algorithms that find resilient and cost-efficient dominating sets.

Economic, social, and evolutionary drivers operate under very different time scales, cost constraints, and objectives. While we have not addressed how the corresponding networks have been developed by their own objectives and cost constraints during decades or hundreds of thousands of years, we showed how drastically different the resulting infrastructure-, social-, and biological networks have become with respect to damage-resilient domination. Among our results we showed that the brain network (extracted from MRI data) is extremely stable against massive failures or attacks; i.e., surviving part of the network (in principle) is structurally capable of carrying out the function of the original network. This finding might provide some insights toward a better understanding of the structural foundations of plasticity. However, this fundamental property comes at a significant cost, capturing the redundancy in the original network (i.e., a high average degree). In contrast, infrastructure networks such as the power grid, whose growth and development is governed by severe cost constraints, are much more vulnerable to failures or attacks in the sense of domination stability.

In the first stage, we have analyzed the effects of damage on various dominating sets based on the *domination stability*, a measure that quantifies the fraction of the network still dominated after certain nodes have been removed. We found that larger dominating sets provide higher stability, and the MDS, which is the most cost-efficient (smallest size) dominating set, is also the most vulnerable, to both random damage and targeted attacks. However, these methods are "fixed" in the sense that they obtain only a single possible dominating set size and corresponding stability for a network. Therefore, our goal was to develop novel methods for finding flexible dominating sets that provide higher stability, and also satisfy cost-efficiency demands.

We have developed two novel methods aiming to overcome the limitations of fixed dominating sets. The first method, the *flexible-redundancy dominating set* (frDS), is an efficient algorithm for obtaining a dominating set with a desired cost (in case of a fixed budget), while maximizing the domination stability. The starting point of our method relies in the observation that the domination stability of the network is reduced by nodes that lose all their dominators. Therefore, we define the *domination redundancy* of a node, denoted by *r*, as the number of dominator nodes it has in its closed neighborhood. Using this measure, we can express the stability locally, as follows. First, we set a desired average domination redundancy *r* in the network, and based on this we assign for each node an *r(i)* domination redundancy requirement, that indicates at least how many dominator neighbors node *i* must have. Then we randomly assign to each node the nearest integer values $\lfloor r \rfloor$ and $\lceil r \rceil$, such that the network average will be *r*. Next, we use a modified greedy algorithm that finds the minimum dominating set, while guaranteeing the preset average domination redundancy. Here, we set the dominating potential *p(i)* as the number of nodes in the closed neighborhood of *i* that have not yet attained their domination requirement. Thus, at each step we select the maximum potential node into the dominating set, until the requirements of all nodes have been satisfied.

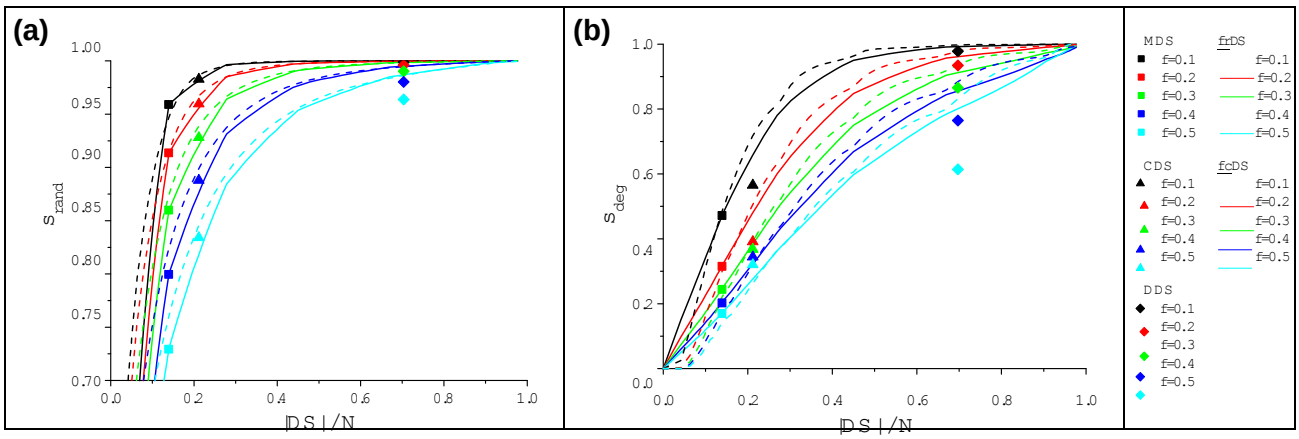


Figure 22. Stability of frDS, fcDS as a function of dominating set size (cost) for various network damage fractions. Subfigure (a) shows random node removal, (b) shows degree-ranked node removal for synthetic scale-free networks, $N = 5000$, $\langle k \rangle = 8, \gamma = 2.5$, averaged over 200 network samples.

Our second method, *flexible-cost dominating set* (fcDS), aims to further optimize the network stability by anticipating attack scenarios, and implementing them in the dominating set construction scheme. For example, if the attack is expected at high degree nodes (as seen in degree-ranked attacks), we should avoid selecting many of those nodes as dominators, despite their ability to cover large fractions of the network. In order to design an algorithm that effectively distributes the dominators in the network, we assign to each node a strength value *s(i)* in the calculation of local stability, which represents the a-priori estimated probability of losing node *i* after the anticipated attack pattern. Next, we calculate the domination stability of node *i* as follows:

$$\text{stability}(\text{DS}, i) = \begin{cases} 0 & \text{if } \text{DS} \cap N^{!}(i) = \emptyset \\ \prod_{!\in!''\cap!'(!)} (1 - s(j)) & \text{otherwise} \end{cases},$$

which is the probability that node *i* will remain dominated. Similarly to the frDS method, we use a modified greedy search algorithm, where at each step we add to the dominating set the highest potential node, in this case the node that increases maximally the stability of the network. Therefore, we define the potential of a node as the total increase of stability it can produce by becoming a dominator node.

$$\text{potential}(i) = \sum_{!\in!'(!)} \text{stability}(\text{DS} \cup \{i\}, j) - \text{stability}(\text{DS}, j) = \sum_{!\in!'(!)} (1 - \text{stability}(\text{DS}, j)) \quad (s(i)).$$

The domination stability of our two flexible strategies frDS and fcDS as a function of redundancy and dominating set size are presented in Fig. 22, where we have also included for baseline comparison the fixed methods (MDS, CDS, DDS), plotted at their corresponding cost values (filled symbols).

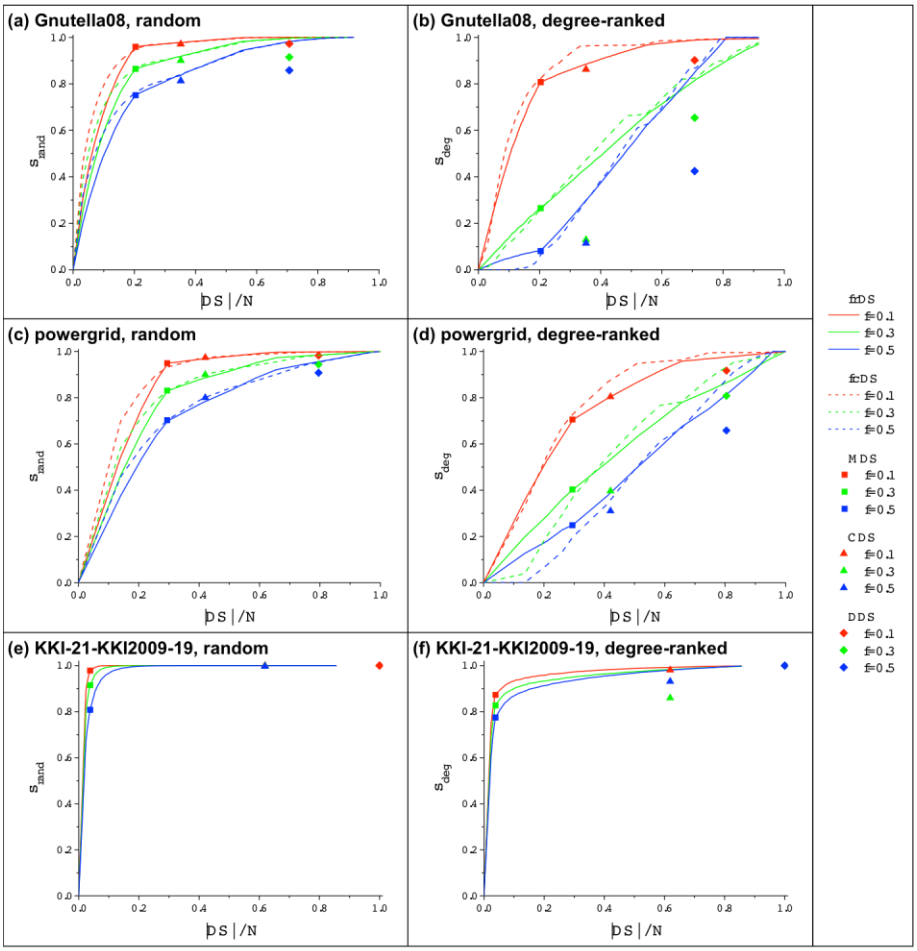


Figure 23. Stability of frDS, fcDS, and other dominating sets in real networks against random and degree-ranked attacks, for various damage fractions: (a,b) Gnutella peer-to-peer network; (c,d) ENTSO-E powergrid; (e,f) Brain (MRI) network. Data is averaged over 20 independent runs of node removal.

Our results show that in case of random damage the stability increases rapidly with cost, and then the curve saturates. Meanwhile, in the degree-ranked damage, there is a steady increase in stability as more nodes are selected as dominators. In both damage scenarios the fcDS provides higher stability than frDS at moderate damage levels, but frDS is more stable at small damage levels. Both frDS and fcDS provide great flexibility in adjusting the size of the dominating set and stability.

In order to validate our results on real-world data, we have studied the domination stability against random failures and targeted attacks on various empirical networks: internet peer-to-peer (Gnutella08) network, the power transmission network of continental Europe (ENTSO-E power-grid) and one brain graph extracted from MRI data (KKI-21-KKI2009-19). Our results are presented in Fig. 23 for both damage scenarios (random failure and targeted attacks). Similarly to the results found for artificial networks, the stability of the frDS and fcDS methods surpasses the stability of CDS and DDS, and matches the stability of MDS, at identical sizes. The stability curves in case of Gnutella08 and power-grid networks saturate slower than seen in artificial scale-free networks. This behavior comes as a result of the non-scale-free degree distribution of these two specific networks. On the other hand, the brain network exhibits high domination stability against both random failure and targeted attack. This particularly high domination stability is the outcome of the peculiar

features of the brain network: high average degree ($\langle k \rangle$ = 138.2) and high assortativity (ρ = 0.62). The high average degree results in highly redundant dominating set, regardless of method, which leads to highly resilient dominating sets in case of a random failure. Meanwhile, the high assortativity ensures that in case of a targeted attack on high degree nodes, the network will have sufficient low degree dominators to protect the domination stability.

We can conclude that the frDS is the most optimal method for finding dominating sets with good stability in case of large-scale networks, or when we have no detailed information about a future attack or damage. On the other hand, the fcDS method can be effectively used to optimize the selected dominating set for high stability, in case we are constrained by a fixed budget for dominating set size, or we have detailed information about potential threats targeting the network.

In summary, we have developed and tested two novel dominating set selection methods that prove to have the highest domination resilience among previously analyzed methods in scale-free networks. Our frDS method with adjustable dominating set size provides efficient protection against random network damage. The fcDS method with adjustable overall resilience, efficiently maximizes the resilience of domination against any potential type of a-priori estimated attack, at any given dominating set size (cost level). We analyzed our methods on a wide range of scale-free network parameters, and performed numerical analysis on multiple real-world network samples, and found similar behavior for both artificial and empirical scale-free networks, demonstrating that our methods can be effectively implemented for practical applications. Our work on the domination stability has been recently published in **[p26]**.

T3. Data-Driven Methods

Efficient Schemes to Evaluate Experiments

There is clear need for statistical evaluation of the experiments in network science. For example, [Centola2010] dicusses the need for experimental design in the social network setting, while [Lu1996] and [Cohn1994] uses experimental designs in engineering applications of network science. In large scale social network data (e.g. Facebook, Twitter, cell-phone network) we have access to a set of parameters, such as gender, level of education, age, race, religion, etc. We want to design statistically sound experiments that allow us to correctly make statistical inferences that answer the following types of questions: What parameter or combination of parameters has the greatest influence on (for example) the number of connections of a person, or on the number of common connections of a connected pair of people have in the social network? Or how does a parameter or a combination of parameters influence the number of connections?

Mixed Orthogonal Arrays (MOAs) are important for statisticians for design of experiments and MOAs of practical size are available on the webpage of the SAS Institute [Kuhfeld]. MOAs reduce the number of experiments needed to draw conclusions. Normally constructions for MOAs depend on finite fields or the existence of other designs. The authoritative text for MOAs is the monograph [Hedayat1997]. Design of experiments is relevant whenever statistical evaluation of the experiments is desired.

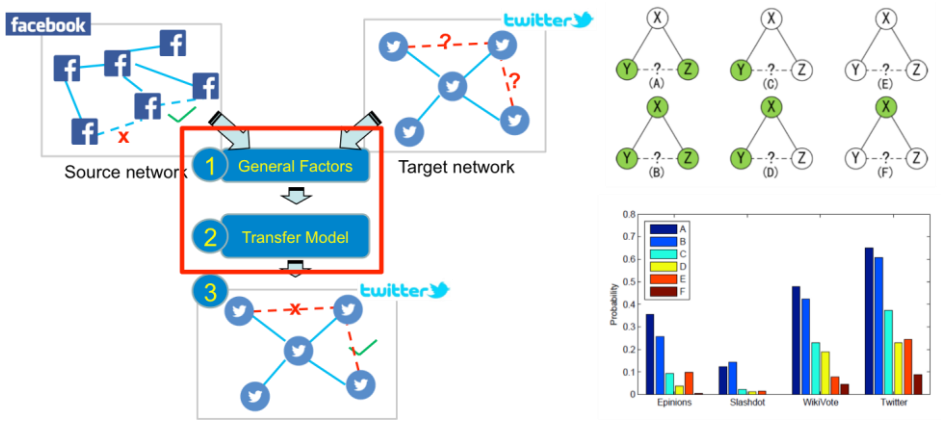
In the reported paper **[p39]**, Aydinian, Czabarka, and Szekely define d-dimensional M-part Sperner families, prove $(n!)$ BLYM inequalities for them (instead of the classical 1) and characterize cases when of all of them hold with equalities, by identifying them with mixed orthogonal arrays. This result settles the issue of connection between Sperner theory and Mixed Orthogonal Arrays.

Extending earlier results [Aydinian2010] for d=1, **[p39]** constructs simple MOAs with constraint M, strength M-d for d>1, basically with rounding, instead of algebraic techniques.
In another work **[41]** Czabarka and Szekely with Matteo Marsili from Italy investigates the following model: A complex system is probed with n repeated experiments, and we only can tell whether two outcomes belong to the same category or not. Consequently outcomes fall into equivalence classes. Assume that k equivalence classes are observed. In 4 random models they worked out threshold functions for n=n(k) as n and k goes to infinity.

Transfer Link Prediction Across Multiple Heterogeneous Social Networks

Link prediction and recommendation is a fundamental problem in social network analysis. The key challenge of link prediction comes from the sparsity of networks due to the strong disproportion of links that they have potential to form to links that do form. Most of the previous work has focused on homogeneous networks. The process of link formation in heterogeneous (or multiplex) networks presents a new set of challenges. In this work, we aim to unveil the interacting human behaviors that underlie the fundamental patterns of social activities. The solution to this problem could help shape and improve our understanding of human behaviors and social networks.

Fig. 24 shows an illustrative example. The top part of Fig. 24a shows two networks—Facebook and Twitter—which is the input of our problem. The bottom part is the output of our problem: formation of new links. The middle of Fig. 24a is the general social patterns we discovered over the two networks for link formation. The fundamental challenge here is how to find the general patterns and bridge them across heterogeneous networks into a unified model for link prediction. We



(a) (b)

Figure 24. Transfer Link Prediction and Recommendation. (a) Transfer Link Prediction Framework; (b) General Factors across Four Social Networks: Preferential Triadic Closure.

formally define the transfer link prediction problem as follows. Given a source network G_S with abundant positive relationships and a target network G_T, the goal is to learn a predictive function f: $(G_T | G_S) \rightarrow Y_T$ for generating the probabilities that a user creates links in the target network by leveraging the information from the source network. This problem formulation is different from the traditional link prediction problem.

We identified various general puspose social factors that may govern the process of link formation in social networks. For instance, we categorize users into two groups (elite users and ordinary users) by estimating the importance of each user by the PageRank algorithm, and selecting the top 1%

users as elite users (opinion leaders), with the others as ordinary users. We try to examine the close triad formation with different types of users in it. The top of Fig. 24b enumerates six cases of the process of triad formation. We examine the probabilities that two users (Y and Z) have a link, conditioned on whether users X, Y, Z are elite users. We find some interesting patterns. First, the probabilities of each of the six cases forming a close triad are very distinct. Second, despite the different sources of the social networks, they share a similar distribution on probabilities of close triad formation in all six cases.

To capture the general factors such as the ones discovered above, we developed a machine-learning model, namely Ranking Factor Graph (RFG), to predict links in a single network and also present a Transfer based RFG (TRFG) model to infer link existence across multiple networks. We first explain the proposed RFG model in detail. We factorize the joint distribution as: $P(Y|G) = \frac{1}{Z}\exp\sum_{i,j,k} f(v_i, v_j, v_k, g(X_i, Y_i))$. This joint distribution contains two kinds of factor functions that may influence the formation of links. The feature correlation factor $f(v_i, v_j, v_k, g(X_i, Y_i))$ represents the influence of an attribute of potential link between v_i and v_j. The social correlation factor $g(X_i, Y_i)$ denotes the influence of social relation Y_i. Finally, we define the log-likelihood objective function as $O(\theta) = \log(p(Y|G)) = \sum_{i,j,k}\sum_{i,j,k}\sum_{i,j,k} \alpha_i f_i(x_{i''i'}, y_{i''}) + \sum_i \beta_i g_i(X_i, Y_i) - \log Z$ where Z is a normalization factor; $|V|$ is the set of users to whom we try to recommend friends and $|C|$ is the candidate list for each user; $\theta = (\{\alpha\}, \{\beta\})$ indicates a parameter configuration. We now turn to discuss how to learn the predictive model with two heterogeneous networks (a source network G_S and a target network G_T. Our intuition is that people make friends in different social networks with similar principles. Back to the model, we use the general patterns (preferential triadic closure) found among different networks and transfer the correlated patterns to help recommend new friends across heterogeneous networks. Straightforwardly, we can define two separate objective functions for source and target networks. The challenge is then how to bridge the two networks such that we can transfer the labeled information from the source network to the target network. Therefore, we define the following log-likelihood objective function over the source and target networks by leveraging general patterns of link formation into the proposed TRFG model: $O(\theta) = \sum_i \alpha_i f_i(.) + \sum_i \mu_i f_i(.) + \sum_i \beta_i(\sum_{i \in i} g_i(X_i, Y_i) + \sum_{i \in i} g_i(X_i, Y_i)) - \log Z$. In this objective function, the first term and the second term respectively define the likelihood over the source network and the target network; the third term defines the likelihood over common features about social patterns defined in the two networks. The common feature functions are defined according to the general social patterns. Such a definition implies that attributes of the two networks can be entirely different as they are optimized with different parameters $\{\alpha\}$ and $\{\mu\}$, while the information transferred from the source network to the target network is the importance of common features that are defined according to the formation of close triads.

The experimental results demonstrate that the proposed RFG model achieves a 10-30% improvement compared with other methods in terms of AUC and achieves a 200% relative improvement compared with unsupervised methods in terms of Precision at Top 30 recommendations. One of the reasons that our RFG has better performance is that it considers some implicit social patterns, namely social balance and close triad formation. In transfer prediction case, we find that TRFG with information transfer from source networks outperforms the RFG without transfer in most cases. We also note that all the transfers from Epinions or Slashdot have a more powerful prediction than RFG. Our results on transfer link prediction have been published in **[p46]**.

Predicting Node Degree Centrality with the Node Prominence Profile

Centrality of a node measures its relative importance within a network. There are a number of applications of centrality, including inferring the influence or success of an individual in a social network, and the resulting social network dynamics. Over the last decade, network evolution modeling focused on defining basic mechanisms driving link creation and capturing macroscopic scaling of real networks. Irrespective of the specific mechanisms that act to drive the emergence of macroscopic scaling, it is reasonable to ask whether such mechanisms also shape the microscopic behaviors of individuals, such as the change in degree centrality. Our findings have important implications for the applications that require prediction of a node's future degree centrality, as well as the understanding of social network dynamics.

We find that the current degree centrality is a weak predictor of the future degree centrality. Rather, the degree centrality evolution is an artifact of both the centrality (preferential attachment) of the node and its relative position (triadic closure) in the network. We define this combination of centrality and position as prominence. To that end, we develop a methodological framework that characterizes the prominence by reconciling the trade-offs between preferential attachment and triadic closure, that is the microscopic level, and develop a model to predict degree centrality of a node in the future. We call our method the Node Prominence Profile (NPP) and was published in **[p31]**.

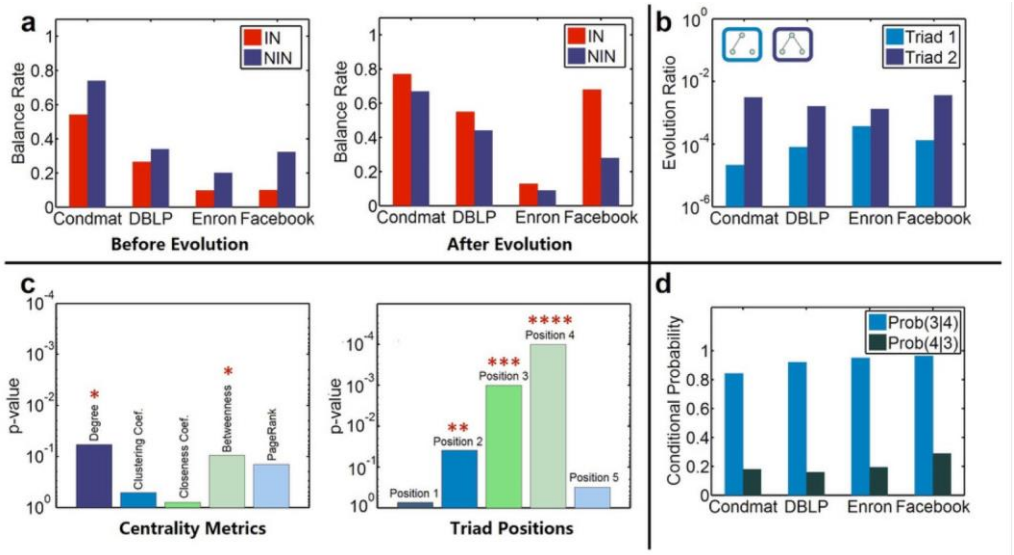


Figure 25. Microscopic Prominence Analysis. (a) Structural Balance Rate. We observe that the balance rates of IN sub-network and NIN sub-network differ before and after the network evolution. (b) Triad Closure Effects. (c) Significance of Inferring Future Degree Centrality. (d) Position Conditional Probability for Position 3 and Position 4.

The empirical experiments reveal that NPP is able to provide more precise prediction of node's future degree centrality over baseline solutions. NPP is validated on four different social networks. We also demonstrate that the model developed on one social network and predict on another social network (transfer learning), thus demonstrating the generalization capacity of NPP and confirming that it is effective in capturing the general factors underlying social network evolution impacting the degree centrality of a node.

NPP optimizes trade-off between essential dimensions of network evolution (preferential attachment and triadic closure). Therefore, it is not surprising that, as a consequence, NPP yields accurate and generic performance in predicting node's future degree centrality. NPP can be effectively used in a variety of applications that rely on inferring a node's importance in the future, as captured by centrality measures. In summary, we have developed a new perspective for predicting the degree centrality of a node in a social network and also developed a general-purpose feature vector that can be used by different machine learning algorithms across different social networks.

Our results on predicting node degree centrality have been published in **[p38].**

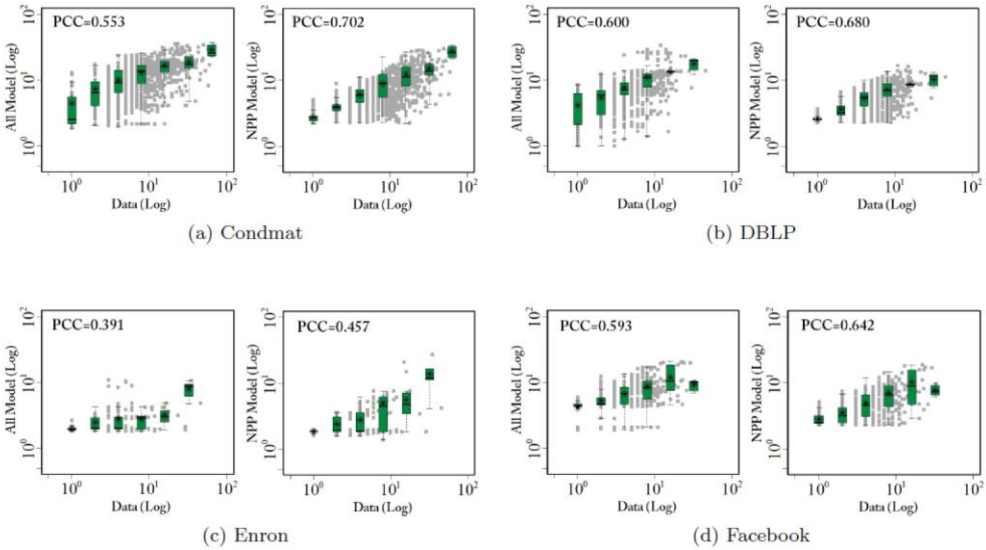


Figure 26. Degree Centrality Prediction Performance. Compare the measured degree centrality (log scale) with the predicted degree centrality (log scale) in four real-world networks. In each subfigure, the left side is the performance of state-of-the-art model and the right side is the performance of NPP model.

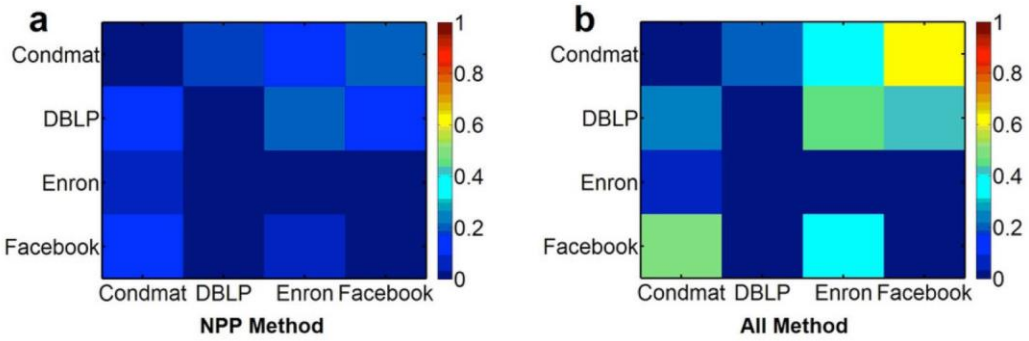


Figure 27. Generalization Performance Loss in AUPR (Degree Centrality Prediction).

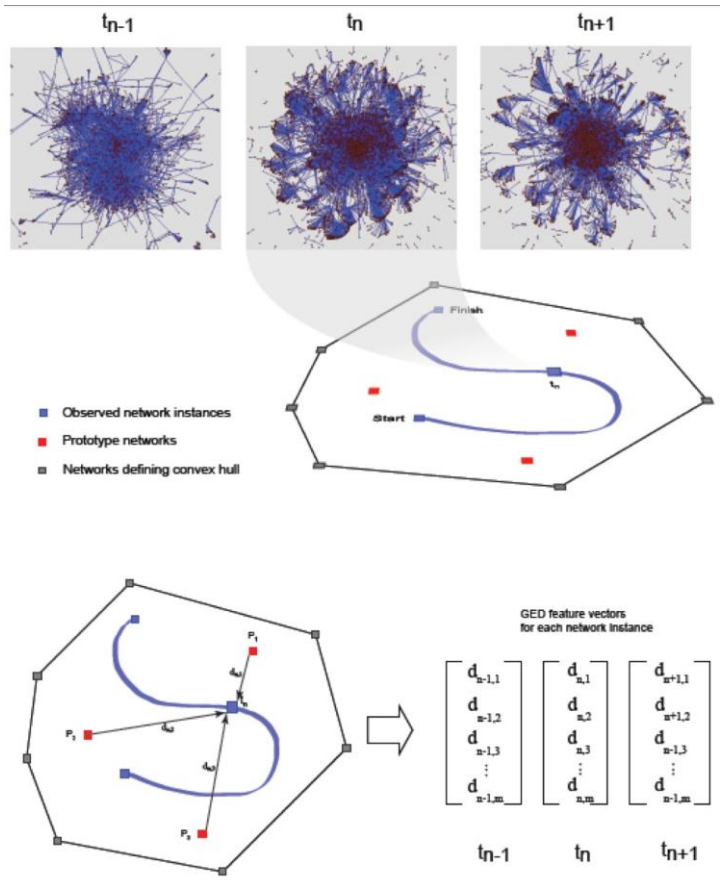


Figure 28. A pictorial summary of the vector space embedding. Networks are extracted at consecutive time windows. The training set defines the convex hull of all possible graphs (as shown in this sketch depicting the MDS of pairwise dissimilarities between graph instances). Prototype networks are extracted, and at each instance, the dissimilarity measure from each of these prototypes represents its GED feature values.

Recurrent Subgraph Prediction (PReSub)

Looking beyond dyadic interactions in dynamic networks:

We present a case for rethinking the way interactions are represented in dynamic networks: instead of considering just dyadic relationships, this concept could be extended to entire subgraphs. To the best of our knowledge, this is a largely unexplored area. In a system of dynamic networks, there exist several recurring subgraphs. The occurrence of these subgraphs can be predicted considering cutting edge link prediction techniques on pairwise edges; another approach would be to consider the subgraph as a whole and predict its occurrence as an entity in its own right. We show that the latter outperforms conventional link prediction on a large swath of parameter sweeps. This suggests that there is an emergent behavior captured by considering subgraphs as the signal, not just their component links. This prediction framework is presented as an out-of-the box pipeline solution (PReSub) to not only predict recurrence of subgraphs in a dynamic network, but also help discover interesting recurring subgraphs which may occur in the system.

Dataset	Number of Nodes	Number of Edges	Time Span
mobile [17]	8,321,119	712M	65 days
wiki [2]	25,323,882	266M	$\sim$ 4 years
enron [24]	87,098	1,147,028	$\sim$ 4 years
facebook [27]	46,715	803,744	$\sim$ 2 years

Table 2. A summary of the data sources used.

PReSub uses the network's vector space embedding as a global set of descriptors and a set of "early warning subgraphs" as local indicators of the subgraph's behavior. Given a labeled temporal network of contact sequences G, one can use PReSub to predict a user-provided subgraph's (S) recurrence, as well as infer recurring subgraphs. A subgraph is considered "recurrent" if it is seen to appear and then subsequently disappear in the training data. For the baseline technique, we took the collection of nodes of the subgraph, i.e. $S.V$ (where $|S.V| = v$), and predicted every link out of the $\binom{v}{2}$ possible links. The subgraph's occurrence is considered to be correctly predicted if and only if its entire linkwise structure is correctly predicted, which means *no other* links are present in the structure. For PReSub, the individual occurrences of the subgraph serve as a supervised ground truth and a predicted occurrence of the subgraph is held to the same strict standard as the baseline method.

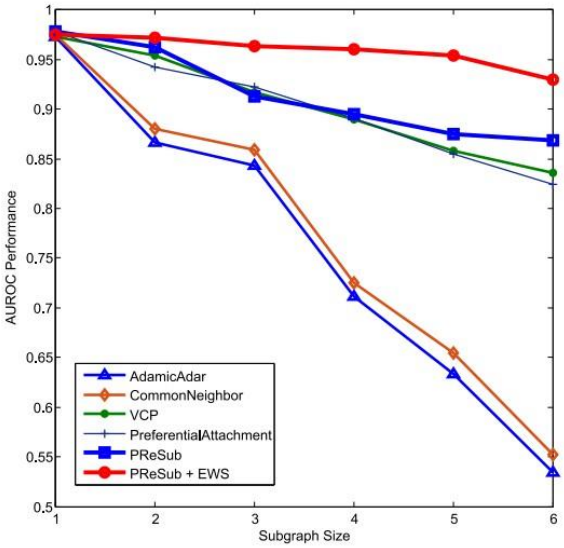


Figure 29. AUROC performance for various methods.

In order to quantify the vector space embedding, we use Graph Edit Distance as a dissimilarity measure, since it is versatile, error tolerant and is capable of dimensionality reduction. Though the closed form computation of it is NP-hard, a near-linear time approximation for computing this measure exists as per. This makes it possible for PReSub to predict recurrent subgraphs in a scalable fashion. This work is currently under review.

We demonstrate our method against various cutting edge link prediction methods implemented in LPMade. We show these to be less effective on area under the ROC curve

(AUROC) than PReSub over a large set of subgraphs on dynamic networks of varying size and duration, as shown in Table 2 above and Fig. 29.

Influence Activation Model: A New Perspective in Social Influence Analysis and Social Network Evolution

Social influence is believed to drive both off-line and on-line human behavior, however it has not been considered as a driver of social network evolution. Our analysis suggests that, while the network structure affects the spread of influence in social networks, the network is in turn also shaped by social influence activity. We posit that the social influence, an intricate mechanism between locality influence and popularity influence, impacts the growth of social networks that may not be captured by the existing network evolution principles. Our goal with this work is to provide a deeper understanding of the social network evolution and develop a model to connect social influence and social network evolution.

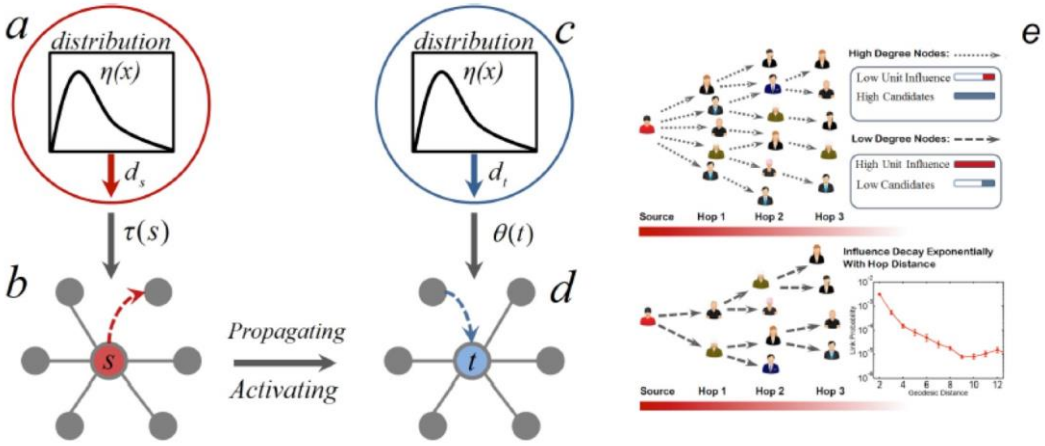


Figure 30. Propagation and Activation Process. In the heuristic of our model, the source node *s* is sending out flux of influence units, each influence unit has an influence activation threshold. In the process of influence unit propagation, it tries to activate each accessible target node. For a target node, it also has a parameter called influence activating ability. When the value of influence activating ability is larger than the activation threshold, the target node is said to be activated by the influence unit from source node.

Consider the notion of Propagation and Activation processes in Fig. 30. Based on the definition of propagation and activation process, we use the number of successful activations between *s* and *t* to estimate the volume of influence between them. We reconcile the locality influence and popularity influence in our model, called the Influence Activation Model (IAM). We consider two aspects for comprehensive evaluation—macroscopic and microscopic validations. At the macroscopic level we consider various network characteristics that define the system or global level properties of a network. Macroscopic study of network focuses on network properties such as degree distributions, diameter, clustering coefficient, geodesic distribution, etc. At the microscopic level, we consider the aspect of link formation in a node's neighborhood, providing a perspective on the nature of human social interactions at a smaller scale to understand the establishment and development of social relationships at a micro-level.

Our work is the first to develop a unified model for network evolution that captures the dynamics of network through the mechanism of influence propagation and link activation driven by social influences. We demonstrate that popularity influence and locality influence are key facets of social influence capital that govern network dynamics. Using different social networks and a variety of macroscopic and microscopic evaluation metrics, we show that IAM is remarkably precise in charting the network evolution. That is, given a current time snapshot of a social network, we can effectively chart the path of network evolution not only from macroscopic network properties but also at the microscopic level to indicate neighborhood activity generated by link formation. These findings and methods are essential to both our understanding of the mechanisms that drive network evolution and our knowledge of the role of social influence in shaping the network structure. Our work on these findings is under submission to Science Advances.

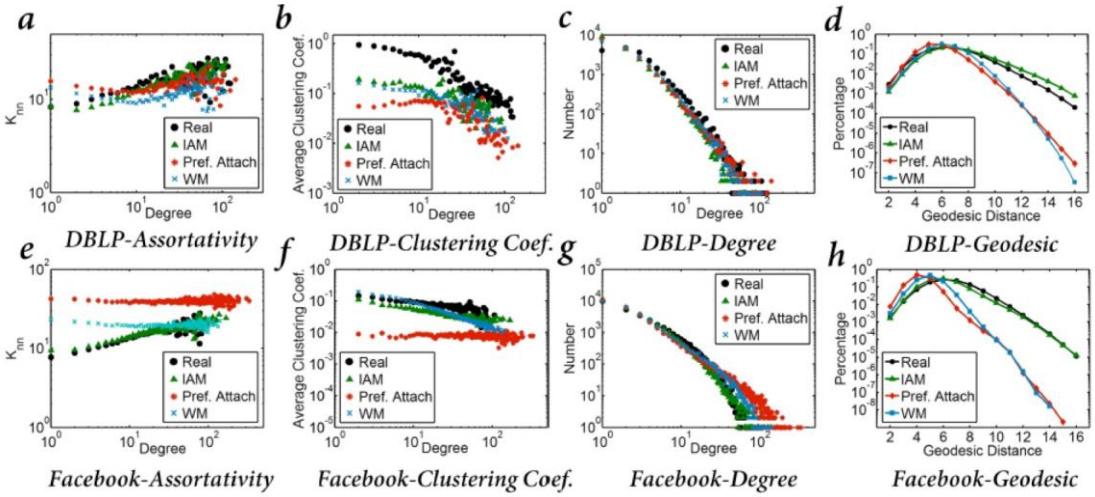


Figure 31. Macroscopic Properties Validation.

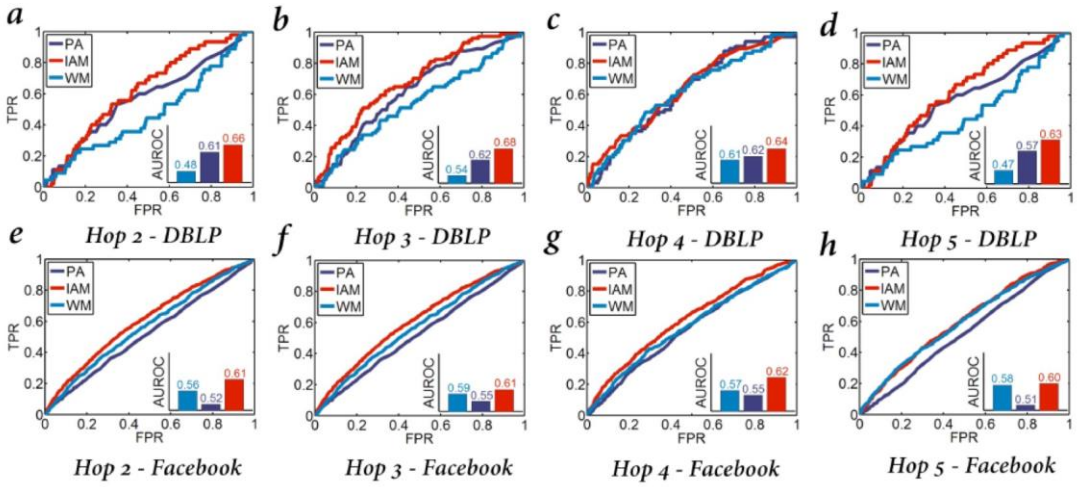


Figure 32. Inferring New Links. Top: The performance (ROC and AUROC) of three models in inferring new links over each geodesic distance (2-5) in DBLP; Bottom: The performance (ROC and AUROC) of three models in inferring new links over different geodesic distance (2-5) in Facebook.

Scientific Impact Modeling and Predicting in Large Academic Social Networks

Scientific impact plays a central role in the evaluation of the output of scholars, departments, and institutions. A widely used measure of scientific impact is citations, with a growing body of literature focused on predicting the number of citations obtained by any given publication. The effectiveness of such predictions, however, is fundamentally limited by the power-law distribution of citations, whereby publications with few citations are extremely common and publications with many citations are extremely rare. Given this limitation, in this work we instead address a related question asked by many academic researchers in the course of writing a paper, namely: "Will this paper increase my *h*-index?" Using a real academic dataset with over 1.7 million authors, 2 million papers, and 8 million citation relationships from the premier online academic service ArnetMiner, we formalize a novel scientific impact prediction problem to examine several factors that can drive a paper to increase the primary author's *h*-index.

We investigate the factors that drive a paper's citation count to become greater than its primary author's *h*-index, including the paper's author(s), content, published venue, and references, as well as social and temporal effects related to its author(s). We examine the importance of different factors as evaluated by correlation coefficients and present the changes of factor importance as predicted for scholars with different *h*-indices (Fig. 33).

According to our correlation analysis, we provide the following intuitions relating to academia: First, A scientific researcher's authority on a topic is the most decisive factor in facilitating an increase in his or her *h*-index. This coincides with the fact that the society fellows (e.g., ACM/IEEE fellow or membership of NAS/NAE) or lifetime honors (e.g., Turing award) are typically awarded for contributions to one specific topic or domain. Second, the level of the venue in which a given paper is published is another crucial factor in determining the probability that it will contribute to its authors' *h*-indices. Top venues make one outstanding and expand one's scientific impact; gradually, one's impact further helps to increase the venue's prestige. Third, people in social society often follow vogue trends. However, publishing on an academically "hot" but unfamiliar topic is unlikely to further one's scientific impact, at least as measured by an increase in one's *h*-index.

We further formalize the question of whether a paper will influence an author's h-index as a novel scientific impact prediction problem. Our prediction task is to determine whether a given paper will, within a pre-defined timeframe, increase the h-index of its primary author (i.e., the

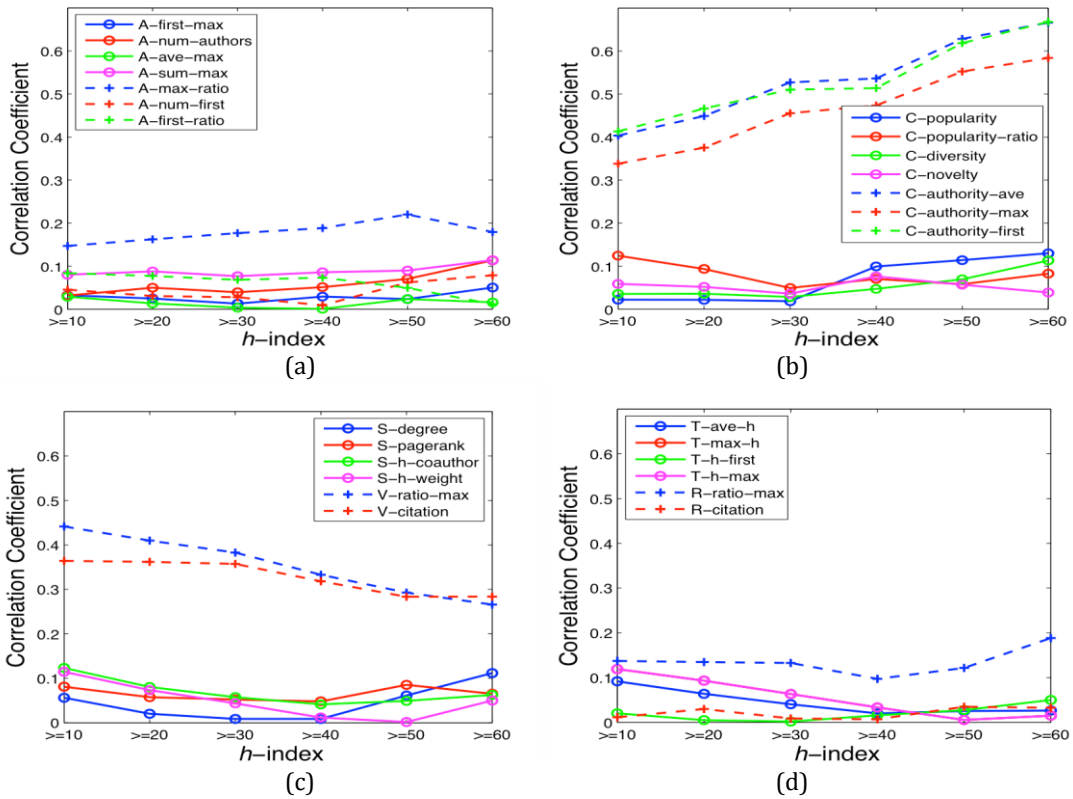


Figure 33. Factor correlation analysis when predicting for scholars with different *h*-indices. *t*=2007 and Δt=5 years. (a) Author factors; (b) Content factors; (c) Social and venue factors; (d) Reference and temporal factors. Author's authority on a subject and published venue are the most highly correlated factors.

researcher with the maximum h-index among the paper's author list). Factors such as the researcher's current influence, the publication topic, and the publication venue may, among many other factors, play a role in determining the degree to which the publication contributes to the researcher's influence. A resulting challenge is the interplay of such factors, which can confound attempts to generate effective predictions. Considerations such as the variability of the h-index according to the "academic age" of a researcher, the widely differing citation conventions among different fields, and the co-authorship of researchers with differing h-indices can make it difficult to isolate the nature and degree to which a given researcher's h-index is influenced by any particular factor. Our work focuses on addressing and overcoming these issues to generate novel, effective scientific impact predictions and to investigate precisely what role a variety of factors play in these predictions.

Given the task of predicting whether a publication will contribute to an author's *h*-index, we find surprisingly strong performance for the problem of scientific impact prediction. Our results demonstrate that we can predict whether a paper will contribute to an author's *h*-index within five years with an F-1 of 0.776. Our study further finds that the most telling factors for determining whether a given paper will contribute to the primary author's *h*-index is the author's authority on the publication topic and the venue in which the paper is published. In contrast, we find that the popularity of the publication topic and the co-authors' *h*-indices are surprisingly inconsequential for determining whether the paper will contribute to the primary author's *h*-index. We also find that (1)

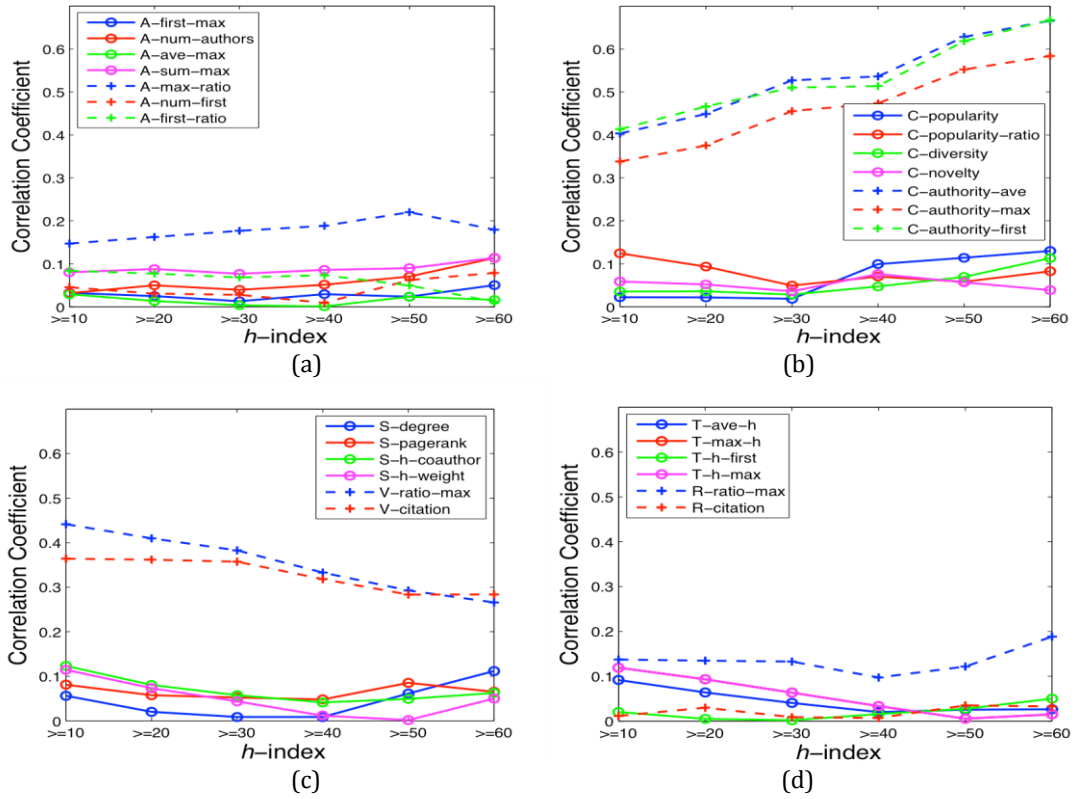


Figure 34. Factor correlation analysis when predicting for scholars with different *h*-indices. t=2007 and Δt=5 years. (a) Author factors; (b) Content factors; (c) Social and venue factors; (d) Reference and temporal factors. Author's authority on a subject and published venue are the most highly correlated factors.

the contribution of papers to researchers with higher *h*-indices is more difficult to predict than it is for researcher's with lower *h*-indices; and (2) the task is more predictable given a long timeframe than a short one.

Overall, our findings unveil mechanisms for quantifying scientific impact and provide concrete suggestions to researchers for better expanding their scientific influence and, ultimately, for more effectively "standing on the shoulders of giants." Notwithstanding the extensive and promising results of the present work, there is still much room left for future work. First, while this work is conducted only on literature from computer science, it is necessary to examine the observed patterns in other scientific disciplines, such as physics, mathematics, biology, and so on. Second, since authors' h-indices evolve within the prediction timeframe, it would be natural to design

methodologies that could capture the co-evolution of authors' h-indices and citation counts. Our results on scientific impact modeling and predicting have been published in **[p25]** (Best Paper Award Nomination 4/39).

Representing Higher Order Dependencies in Networks

A network is a representation of data or events, and there can be various network representations from data, depending on how edges are defined and/or weighted. Whether a network is an accurate representation of the data is the prerequisite for meaningful network analyses. While a common norm is to build the network by leveraging pair-wise or dyadic interactions between nodes, it implicitly assumes Markov property (first order dependency). For data with higher order dependencies that exist ubiquitously in real-life, patterns can go beyond dyadic or pairwise

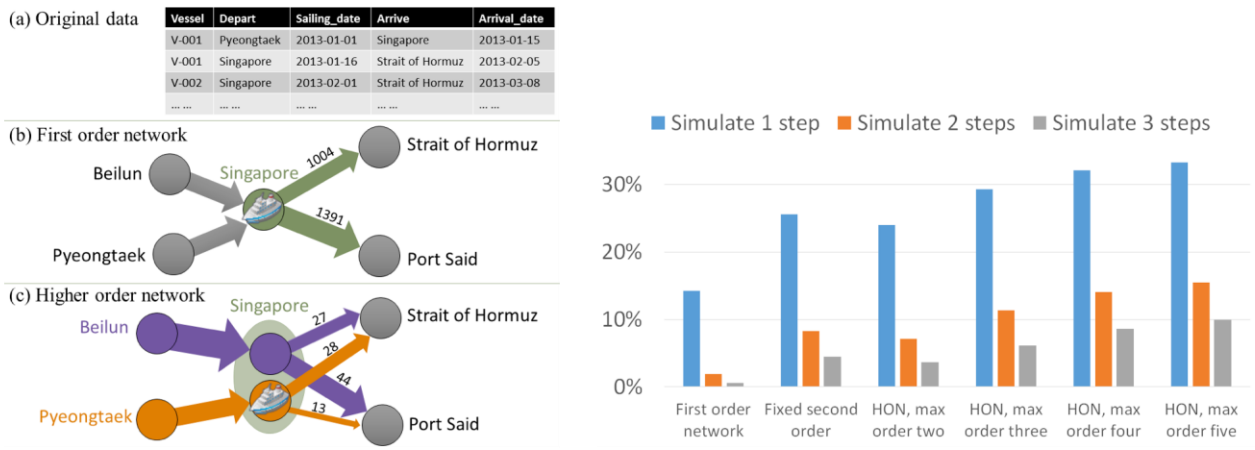


Figure 35. Different ways of constructing network from data. Left: (a) shows an example of global shipping data as a list of recorded events. (b) shows the conventional way of constructing data where every node represents a port, every edge representing the annual shipping traffic between ports. This network can only model first order dependencies and cannot capture the fact that vessels coming from different ports to Singapore will have different probability distributions for the next step. (c) shows our proposed network representation where higher order dependencies in data are effectively modeled in the network representation. Right: using our proposed High Order Network (HON) instead of conventional first order network or fixed second order network, higher order dependencies in data are preserved in our network representation, so that the simulation of movements on the network will see significant improvements on the accuracies.

interactions, so the conventional way of network construction oversimplifies the connections in data and consequently lead to inaccuracies in network analysis, particularly when applying a wide range of network analysis tools that are based on the simulation of movements on the network, such as clustering with MapEquation, Ranking with PageRank, various link prediction methods, and so on.

We presented an algorithm to construct network from data with higher order dependencies. Instead of assuming a fixed order for the whole network, we extracted the orders of dependencies for every path from the data, then embedded these higher order dependencies into a network via a compact representation. Our algorithm is flexible allowing for variable order of dependencies in the data. Our proposed High Order Network (HON) is (1) able to represent higher order dependencies in data, so that the simulation of movements on the network can follow higher order dependency patterns and

be more accurate; (2) compact in size by adding nodes and edges to a first order network only where necessary, and (3) compatible with all existing network analysis tools, so that even if tools based on simulation of movements are unchanged, by using HON instead of first order network, these tools can benefit from the improved simulation of movements.

With four data sets, we showed our proposed method is effective in extracting higher order dependencies from data, and our proposed network representation HON can effectively represent these higher order dependencies towards higher accuracies of movement simulation on networks. For global shipping data, the accuracy for simulating one step on HON has more than doubled compared with first order network, and higher by one magnitude when simulating multiple steps.

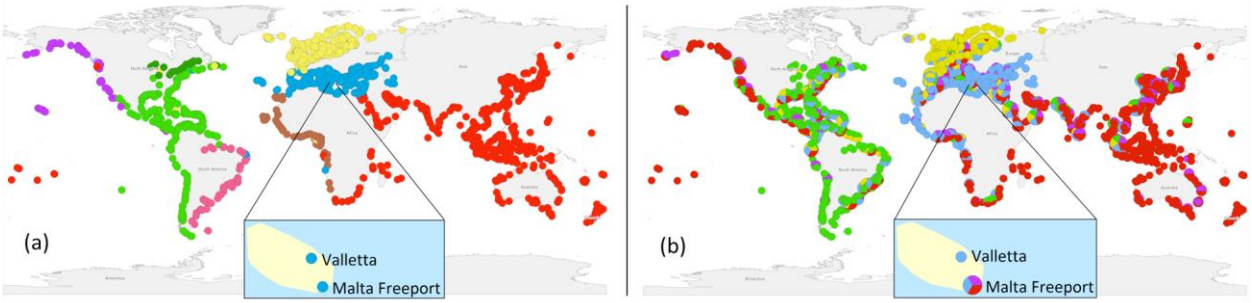


Figure 36. Clustering of ports on different network representations of global shipping data. **(a)** is the clustering on first order network where non-overlapping clusters are generated; **(b)** is on HON where clusters may overlap and international ports (such as Malta Freeport) are effectively identified by belonging to multiple clusters.

For network analysis methods based on simulation of movements, we also show that by using HON instead of first order networks those methods can yields more insights (overlapping clusters for MapEquation indicating international ports in global shipping data, and changes of rank for PageRank promoting / depressing web pages). Because there have been lots of network analyses based on first order networks, we expect to see positive changes when using HON instead to account for higher order dependencies in data. A wide range of applications such as invasive species analysis based on clustering of global shipping network, advertisement system based on ranking of web pages, urban planning based on simulation of traffic, can potentially be influenced. In future work we look forward to extending the application of HON beyond simulation of movements to more dynamic processes, and improve the algorithm by reducing the parameters needed. This work is under review at KDD 2015.

References:

[Adcock2013] A. Adcock, B. D. Sullivan, and M. Mahoney. Tree-like structure in social and information networks. *Proceedings of 2013 IEEE International Conference on Data Mining (ICDM'13)* (2013).

[Adcock2014] A. Adcock, B. D. Sullivan, and M. Mahoney. Tree decompositions and social graphs. Available on arXiv (2014).

[Alon2000] N. Alon, J. H. Spencer. *The Probabilistic Method, second edition*. John Wiley and Sons (2000).

[Aydinian2010] H. Aydinian, E. Czabarka, K. Engel, P. L. Erdos, L. A. Szekely, A note on full transversals and mixed orthogonal arrays. *Australasian J. Comb.* **48**, 133-141 (2010).

[Barefoot1997] C.A. Barefoot, R.C. Entringer, L.A. Szekely. Extremal values for ratios of distances in trees. *Discrete Appl. Math.* **80**, 37-56 (1997).

[Centola2010] D. Centola, The Spread of Behavior in an Online Social Network Experiment. *Science* **329**, 1194-1197 (2010).

[Cohn1994] D.A. Cohn, Neural network exploration using optimal experiment design, A.I. Memo No. 1491, MIT Artificial Intelligence Laboratory (1994).

[Cooper2007] C. Cooper, M. Dyer and C. Greenhill. *Comp. Prob. Comp.* **16**(4), 557-593 (2007).

[Dehmer2008] M. Dehmer. A Novel Method for Measuring the Structural Information Content of Networks. *Cybernet. Syst.* **39**, 825-843 (2008).

[Dehmer2011] M. Dehmer, F. Emmert-Streib, Y. Tsoy, and K. Varmuza. Quantifying structural complexity of graphs: Information measures in mathematical chemistry. *in: Quantum Frontiers of Atoms and Molecules*, 479-498 (2011).

[Erdos2010] P.L. Erdos, I. Miklos, Z. Toroczkai. *Elec. J. Comb.* **17**(1), R66 (2010).

[Genio2010] C.I. Del Genio, H. Kim, Z. Toroczkai and K.E. Bassler. PLoS ONE **5**(4), e10012 (2010).

[Gleiser2003] P. Gleiser, L. Danon. Community structure in jazz. *Advances in Complex Systems* **06**(04), 565–573 (2003).

[Greenhill2011] C. Greenhill. *Electronic J. Comb.* **16**(4), 557-593 (2011).

[Handcock2003] M.S. Handcock, G.L. Robins, T.A.B. Snijders. Assessing degeneracy in statistical models of social networks, Center for Statistics and the Social Sciences Working Paper no. 39 (2003).

[Hedayat1997] A. S. Hedayat, N. J. A. Sloane, J. Stufken. *Orthogonal Arrays: Theory and Applications, Springer Series in Statistics*. Springer (1999)

[Jaynes1957] E.T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review* **106**(4), 620–630 (1957).

[Jerrum1986] M.R. Jerrum, L.G. Valiant and V.V. *Vazirani. Theor. Comput. Sci.* **43**, 169 (1986).

[Kannan1999] R. Kannan, P. Tetali and S. Vempala. *Rand. Struct. Alg.* **14**(4), 293-308 (1999).

[Kim2009] H. Kim, Z. Toroczkai, P.L. Erdos, I. Miklos, L.A. Szekely. J. Phys. Math. Theor. **42**, 392001 (2009).

[Kim2012] H. Kim, C.I. del Genio, K.E. Bassler and Z. Toroczkai. *New J. Phys.* **14**, 023012 (2012).

[Kuhfeld] W.F. Kuhfeld. Orthogonal Arrays (SAS Tech Note TS-723), http://support.sas.com/techsup/technote/ts723.html

[Liljeros 2001] F. Liljeros et al., The web of human sexual contacts. *Nature* **411**(6840), 907–908 (2001).

[Lu1996] C-L Lu, T.K. Fong, R.T. Hofmeister, P. Poggiolini, L.G. Kazovsky. CORD-A WDM optical network: design and experiment of fast data synchronization by pilot-tone transport. *Photonics Technology Letters* **8**(8), 1070-1072 (1996).

[Lu2007] L. Lu and L.A. Szekely. Using Lovasz Local Lemma in the space of random injections. *El. J. Comb.* **14**, R63 (2007).

[Miklos2012] I. Miklos, E. Tannier. Approximating the number of double cut-and-joinscenarios. *Theoretical Computer Science* **439**, 30-40 (2012).

[Miklos2014] I. Miklos, S. Kiss, E. Tannier. On sampling SCJ rearrangement scenarios. *Theoretical Computer Science* **552**(2), 83-98 (2014).

[Moser2010] R. Moser, G. Tardos. A constructive proof of the general Lovász Local Lemma. *Journal of the ACM***, 57**(2), Article No. 11 (2010)

[Nacher2012] J.C. Nacher et al., *New Journal of Physics* **14**, 073005 (2012)

[Palmer1988] J.D. Palmer. Intraspecific variation and multicircularity in Brassica mitochondrial DNAs. *Genetics* **118**, 341-351 (1988).

[Park2004] J. Park, M.E.J. Newman. Solution of the two-star model of a network. *Physical Review E* **70**(6), 066146 (2004).

[Park2005] J. Park, M.E.J. Newman. Solution for the properties of a clustered network. *Physical Review E* **72**(2), 026136 (2005).

[Robins2007] G.L. Robins et al., An introduction to exponential random graph (p*) models for social networks. *Social Networks* **29**(2), 173–191 (2007).

[Snijders2006] T.A.B. Snijders et al., New Specifications for Exponential Random Graph Models. *Sociological Methodology* **36**(1), 99–153 (2006).

[Stanton2012] I. Stanton, A. Pinar. Constructing and sampling graphs with a prescribed joint degree distribution. *ACM Journal on Experimental Algorithms* **17**, Article No. 3.5 (2012).

[Sun2009] Y. Sun, B. Danila, K. Josić, K. E. Bassler. Improved community structure detection using a modified fine-tuning strategy. *EPL* **86** 28004 (2009).

[Szekely2005] L.A. Szekely, H. Wang. On subtrees of trees. *Adv. Appl. Math.* **34**, 138-155 (2005).

[Szekely2006] L.A. Szekely, H. Wang. Binary trees with the largest number of subtrees. *Discrete Appl. Math.* **155**(3), 374-385 (2006).

[Vazirani2003] V.V. Vazirani. *Approximation Algorithms*. Springer (2003).

[Wagner07] S. Wagner. Correlation of graph-theoretical indices. *SIAM J. Discrete Mathematics* **21**(1), 33-46 (2007).

[Zachary1977] W. Zachary. An information flow model for conflict and fission in small groups. *Journal of Anthropological Research* **33**(4), 452–473 (1977).

1.

1. Report Type

Final Report

Primary Contact E-mail

Contact email if there is a problem with the report.

toro@nd.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

574-631-2618

Organization / Institution name

University of Notre Dame

Grant/Contract Title

The full title of the funded effort.

DARPA Ensemble-Based Modeling Large Graphs & Applications to Social Networks

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-12-1-0405

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Zoltan Toroczkai

Program Manager

The AFOSR Program Manager currently assigned to the award

James Lawton

Reporting Period Start Date

09/01/2012

Reporting Period End Date

07/31/2015

Abstract

Modeling and predicting the structure and behavior of real-world complex networks, including those with defense applications requires a suite of computational and mathematical tools that need to address model generation, network search and data analysis problems. During this project we have achieved our main proposed goals in all these research fronts. T1. Modeling with Constrained Graph Ensembles. We provided results to characterize graph ensembles defined by empirical data in form of sharp constraints, including existence, construction, sampling and graph counting problems. Developed novel proofs for the Markov Chain Monte Carlo mixing time problem for graph sampling based on degrees and joint degrees. For Exponential Random Graph Models we provided an understanding of the degeneracy problem and proposed a novel method that eliminates this problem. Described spectral properties of random geometric graph ensembles in hyperbolic spaces and connected them to real social networks. T2. Finding Structures of Interest. Shown that minimum dominating sets (MDS) provide an effective way to search, monitor and influence large networks. Developed linearly scalable heuristic algorithms based on the probabilistic method (Lovasz Local Lemma) to identify MDS in large networks; we demonstrated its applicability on a model for opinion spread. Expanded the setting for the applicability of LLL for network problems to be also used as a counting tool. T3. Data Driven Methods. Developed novel results for Mixed Orthogonal

Arrays to be used for the design of experiments and statistical queries of large networks. Developed novel tools for link prediction, node prominence prediction, and influence propagation in multirelational networks. Modeled co-evolutionary processes for subgraph – embedding graph relationships.

DoD Relevance: Economies, social and political systems all exist embedded within complex and dynamic networks.
Understanding the network landscape is a critical necessity for winning conflicts and securing global safety, peace and stability. It helps prevent surprises and provides both a tactical and a strategic upper hand in our interactions both with our allies and our adversaries. Understanding complex networks requires the development of reliable mathematical and computational tools that efficiently probe the data and extract the relevant information. This is a very difficult undertaking and it requires marshaling methods from traditionally disparate areas, including graph theory (discrete mathematics), statistical physics, statistics, theoretical computer science, algorithm development, and data mining in a sustained fashion. This project has been developing mathematical and computational methods pushing the state of the art in data driven modeling of complex networks. The algorithms and the methods developed here have been validated against real-world network datasets.

Output: The output from this research is in form of publications, preprints, presentations at conferences and meetings. A total of 47 publications and preprints with another 13 publications under preparation, 7 PhDs and 1 MSc theses generated. Other highlights include students graduated (8), faculty awards (2), conference presentations (69) and other activities and interactions (such as conferences organized), see the sections below.

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

AFD-070820-035.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

AFOSR_NDFinal_Report.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

A total of 47 publications were generated in peer reviewed journals, many in the high impact and prestigious journals. Please see the full report for details.

Changes in research objectives (if any):

N/A

Change in AFOSR Program Manager, if any:

Previously: Bob Bonneau.

Extensions granted or milestones slipped, if any:

There was a descoping on this project from DARPA side. No milestones have sliped, but results were accelerated and finalized earlier.

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Jul 29, 2015 02:18:01 Success: Email Sent to: toro@nd.edu