SIAM Workshop on Network Science 2018 (NS18)

July 12 - 13, 2018, Portland, OR, USA

List of Abstracts
(in order of presentation)

Invited Talk I

Percolation and cascades on networks

Raissa M. D'Souza

University of California, Davis

Mathematical models of the structure and dynamics of networks provide theoretical tools for studying the complex and interdependent networks of modern society. Two central themes of these models are percolation, the emergence of large-scale connectivity in an a network, and cascading failures. This talk will survey how we use the paradigms of random graphs, rate equations and non-linear dynamics to address these issues. We begin by considering the structure of a network and the novel classes of percolation phase transitions that result from repeated, small interventions intended to delay the onset of connectivity. We then move on to considering dynamics on a network, in particular cascading failures on coupled networks and the notions of optimal interdependence, architectures that minimize cascades, and catastrophe-hopping leading to non-local cascades. The talk concludes with some discussion of future research directions.

HETEROGENEOUS MICRO-STRUCTURE OF PERCOLATION ON NETWORKS

Tim Rogers

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

What is the probability that a particular node appears in the giant component of a network in a random instance of supercritical bond percolation? If it is not in the giant component, then what is the expected size of its component? Perhaps surprisingly, there is in fact considerable variation between nodes in both of these quantities. In this talk I will unpack this heterogeneous micro-structure of percolation for sparse networks using message-passing, population dynamics, and weakly non-linear analysis.

Research Questions

Since the very begining of the modern fascination with networked systems, researchers have been interested in questions of propagation. Across many applications, the analysis of bond percolation provides a simple framework with which to analyze the capability of a network to transmit information, disease, influence, or failure [2, 1, 8, 10]. Early work in this area mainly concentrated on understanding the global properties of percolation in the ensemble average of randomly generated model networks. Surprisingly, detailed results for single instances of fixed networks have only been available relatively recently [5, 3], and very little is known exactly about the responses of individual nodes [9].

Consider bond percolation for fixed networks defined as follows: starting from an arbitrary large (connected) network, we evaluate each edge independently, keeping it with probability ρ and deleting it with probability $1-\rho$. The largest connected component remaining after this random edge removal process is referred to as the *percolating cluster* or *giant component*; write S for its size measured as a fraction of the total number N of nodes in the network. For large sparse networks it was shown in [4, 5] that this quantity can be computed to close approximation using a message-passing protocol.

In this talk (based on joint work with Reimer Kühn [6]) I will ask more detailed questions about the typical outcomes for individual nodes in the network, when averaged over many instances of the percolation process. For

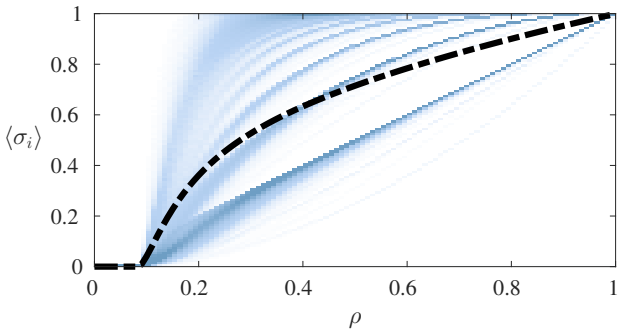


Figure 1: Micro-structure of percolation in a sample network with 62,586 nodes taken from the *gnutella* file sharing platform [7]. Each vertical slice of the density plot shows the distribution $\varphi(s)$ of probability to appear in the percolating cluster for given edge occupation probability ρ. The thick dashed line shows the expected size of the percolating cluster S, which is equal to the mean of φ.

a given random instantiation of percolation, write $\sigma_i = 1$ if node i appears in the largest connected component, and $\sigma_i = 0$ if not. Taking the ensemble average of this variable yields the probability $\langle\sigma_i\rangle$ for node i to appear in the percolating cluster. Heterogeneity in the responses of individual nodes to percolation is captured by the empirical distribution of $\langle\sigma_i\rangle$, defined as

$$\varphi(s) = \frac{1}{N}\sum_i \delta(s - \langle\sigma_i\rangle)\,. \tag{1}$$

The total fractional size of the percolating cluster is given by the mean of φ, that is, $S = \int s\,\varphi(s)\,\mathrm{d}s$. However, for many networks, this average is not at all representative of the behaviour of individual nodes: see Figure 1 for an example. This heterogeneity was observed in the context of epidemic spread in [9], which for the analysis of SIR dynamics allows an exact mapping on bond percolation. In that work it was found that node degrees play a dominant role in the behaviour of $\langle\sigma_i\rangle$ near $\rho = 1$, but also that the picture becomes much more complex near criticality.

When node i does not appear in the percolating cluster, write n_i for the size of the component it belongs to, and

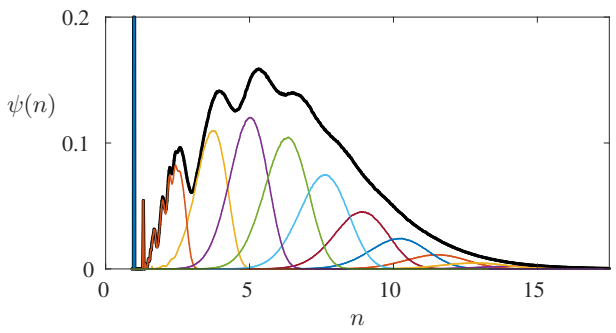


Figure 2: Distribution $\psi(n)$ of average cluster sizes for percolation on an Erdős-Rényi network of mean degree $c = 4$ at $\rho = 0.3$. Results of population dynamics shown together with its unfolding according to degree for $k = 0, 1, 2, \ldots, 9$, and $\{k \geq 10\}$ (blue, red, green,... from left to right).

$\langle n_i \rangle$ for the average over many instances. The node average $\frac{1}{N}\sum_i \langle n_i \rangle$ was again analysed for finte networks in [5], and previously results for the distribution of finite cluster sizes in large random graphs was presented in [1]. The empirical distribution small clusters is defined analogously to φ in Eq. (1);

$$\psi(n) = \frac{1}{N}\sum_i \delta(n - \langle n_i \rangle)\,. \tag{2}$$

Just as for the probability to appear in the percolating cluster, we also observe a broad distribution for the sizes of small clusters associated to different nodes. Figure 2 shows an example of this distribution for Erdős-Rényi networks, disaggregated according to node degree.

Methodology

As detailed previously in [5], analysis of the probability generating function of component sizes yields a set of self-consistency equations which can be solved efficiently by iteration. For a network with M edges we define the $2M$-vector $\boldsymbol{H}$ to be the smallest solution in $[0, 1]$ of the system

$$H_{i\leftarrow j} = (1-\rho) + \rho \prod_{\ell \in \mathcal{N}_j \setminus i} H_{j\leftarrow \ell}\,. \tag{3}$$

Here we write $\mathcal{N}_j$ for the neighbourhood of node j, and the entries of the vector $\boldsymbol{H}$ are indexed by ordered pairs of nodes attached by an edge. The probability of node i to appear in the percolating cluster is recovered via

$$\langle \sigma_i \rangle = 1 - \prod_{j \in \mathcal{N}_i} H_{i\leftarrow j}\,. \tag{4}$$

Linearizing the system (3) around the state $\boldsymbol{H} = \mathbf{1}$ yeilds the so-called *non-backtracking* matrix B with entries $B_{i\to j, k\to \ell} = 1$ if $j = k$ and $i \neq \ell$, and zero otherwise. It is known that the top eigenvalue of this matrix gives the reciprocal of the critical edge density ρ_c [5, 3]. As I will explain in this talk, weakly non-linear analysis near criticality allows $\langle \sigma_i \rangle$ and $\langle n_i \rangle$ to both be recovered from the corresponding top eigenvector.

Moreover, in the limit of large networks (e.g. from the Erdős-Rényi ensemble), it is possible to draw numerical samples from the distributions φ and ψ by iterating the message passing equations (3) with random topologies — this population dynamics technique was used to generate Figure 2. Further, I will show that if the mean degree is large, then analysis of the population dynamics equations leads to the following asymptotic forms for the distributions of percolation probability and finite cluster size:

$$\varphi(s) \approx \frac{\exp\left\{-\frac{c}{2}\left(1 + \frac{1}{\eta}\log(1-s)\right)^2 - \log(1-s)\right\}}{\sqrt{2\pi/c}\,\eta}\,.$$

$$\psi(n) \approx \frac{\sqrt{c}}{\sqrt{2\pi}\gamma}\exp\left\{-\frac{c}{2\gamma^2}\left(n - 1 - \gamma\right)^2\right\}\,.$$

References

[1] D. S. Callaway, M. E. Newman, S. H. Strogatz, and D. J. Watts. Network robustness and fragility: Percolation on random graphs. *Physical review letters*, 85(25):5468, 2000.

[2] P. Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63(2):157–172, 1983.

[3] K. E. Hamilton and L. P. Pryadko. Tight lower bound for percolation threshold on an infinite graph. *Phys. Rev. Lett.*, 113:208701, 2014.

[4] B. Karrer and M. Newman. Message passing approach for general epidemic models. *Physical Review E*, 82(1):016101, 2010.

[5] B. Karrer, M. E. J. Newman, and L. Zdeborová. Percolation on sparse networks. *Phys. Rev. Lett.*, 113:208702, 2014.

[6] R. Kühn and T. Rogers. Heterogeneous micro-structure of percolation in sparse networks. *EPL (Europhysics Letters)*, 118(6):68003, 2017.

[7] J. Leskovec and A. Krevl. SNAP Datasets: Stanford large network dataset collection. `http://snap.stanford.edu/data`, 2014.

[8] M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[9] T. Rogers. Assessing node risk and vulnerability epidemics on networks. *Europhysics Letters*, 109:28005, 2015.

[10] D. J. Watts. A simple model of global cascades on random networks. *Proceedings of the National Academy of Sciences*, 99(9):5766–5771, 2002.

A RE-ENTRANT PHASE TRANSITION IN THE SURVIVAL OF SECONDARY INFECTIONS ON NETWORKS

Sam Moore, Peter Mörters, Tim Rogers

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We study the dynamics of secondary infections on networks, in which only the individuals currently carrying a certain primary infection are susceptible to the secondary infection. In the limit of large sparse networks, the model is mapped to a branching process spreading in a random time-sensitive environment, determined by the dynamics of the underlying primary infection. When both epidemics follow the Susceptible-Infective-Recovered model, we show that in order to survive, it is necessary for the secondary infection to evolve on a timescale that is closely matched to that of the primary infection on which it depends.

Abstract

Superinfections are a major cause of global mortality and morbidity. For example, the WHO estimates 15 million cases worldwide of Hepatitis D, which spreads only amongst carriers of Hepatitis B and greatly worsens their prognosis [1]. There is a need, therefore, to develop a robust understanding of the conditions under which outbreaks of secondary infections are possible.

In 2013, Court, Blythe and Allen [3] introduced a model of hierarchical infection referred to as the *stacked contact process*. Their model concerns the fate of a population of coevolving hosts, spreading as a contact process on a lattice, and parasites, spreading as a contact process restricted to sites currently occupied by hosts. In epidemiological language, the contact processes of [3] correspond to coupled Susceptible-Infective-Susceptible (SIS) epidemics; empty lattice sites are interpreted as susceptible individuals, who may be infected by the primary (host) and then secondary (parasite) infections.

At around the same time, Newman and Ferrario [7] independently proposed a related model in the context of epidemic dynamics in social contact networks. They considered a pair of Susceptible-Infective-Recovered (SIR) epidemics with a strictly obligate relationship such that the secondary infection is only transmitted amongst those who have *recovered* from the primary. In this formulation the dynamics of the two diseases are completely separated in time, allowing for analytical treatment of the model using "cavity method" techniques which have been quite successful in the study of epidemics on networks (see, e.g. [6, 8]).

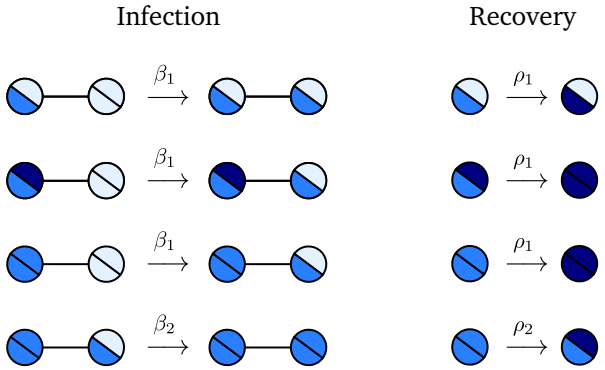


Figure 1: Possible events and their rates in the network superinfection model. Circles represent nodes in the network, with the state of the primary (resp. secondary) infection shown by the colour of the lower-left (resp. top-right) sector; light denotes susceptible, midtone denotes infective, dark denotes recovered.

The introduction of network structure to the population in [7] has the advantage of improving the relevance of the model for human epidemic dynamics, however, by separating the dynamics of the two diseases this model is limited in its description of the interaction between infection timescales, characterisation of which form the main results in Court, Blythe and Allen's model [3, 5].

We then study the dynamics of coevolving SIR superinfections in sparse contact networks by considering a population of individuals occupying the vertices of a Erdős-Rényi random graph with mean degree c. A primary infection spreads through the population with infective individuals passing the disease on to their neighbours with rate β_1, and recovering from the disease with rate ρ_1. Individuals who are carrying a *live* primary infection may also play host to a secondary infection, which spreads and recovers with rates β_2, ρ_2. See Fig. 1 for an illustration of the possible state transitions.

The success of the primary infection is controlled by the connectivity (mean node degree) c of the network, and the ratio

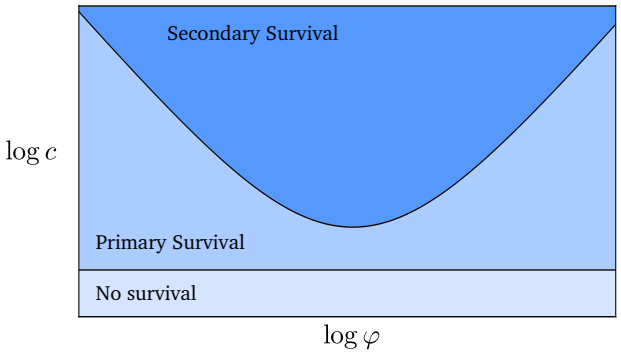


Figure 2: Schematic phase diagram of the superinfection network model for fixed $\beta_1/\rho_1 = \beta_2/\rho_2$, shown as a function of the relative timescale $\varphi = \beta_1/\beta_2$ of the infections and the connectivity c of the network. The secondary infection survives with positive probability only in a convex region whose boundary is characterised in our Theorem 1.

of the infection and recovery rates $\alpha := \beta_1/\rho_1$. This parameter is well understood as the basis of the classical single infection process: for fixed α there exists a critical value of c above which the infection survives within finite probability and at or below which we have certain extinction, see e.g. [4].

Concentrating on the case that $\beta_2/\rho_2 = \alpha$ also, three parameters then describe success of secondary infection: the connectivity of the underlying graph, c; the ratio between spread and recovery, α; and, crucially, the relative timescales of the two infections, $\varphi := \beta_1/\beta_2$.

In order for the secondary infection to survive it is perhaps intuitive that it must progress at a rate fast enough compared to the primary infection, else the primary infection will have itself recovered and subsequently ended the secondary infection before the secondary infection has a chance to spread. Perhaps more surprisingly however we also show that the secondary infection should not act too quickly as this too compromises survival potential.

Our characterisation of the survival of the secondary infection is illustrated in Fig. 2 and summarised by our main result:

Theorem 1 *For all $\alpha, \varphi > 0$ there exists a critical connectivity $c^\star$ such that, in the limit of large network size, for $c < c^\star$ the secondary infection dies out with probability one, and for $c > c^\star$ it survives with positive probability.*

Furthermore, for large and small φ we have the scaling behaviour

$$c^\star \sim \varphi \quad \textit{for} \quad \varphi \to \infty\,, \qquad c^\star \sim 1/\varphi \quad \textit{for} \quad \varphi \to 0\,.$$

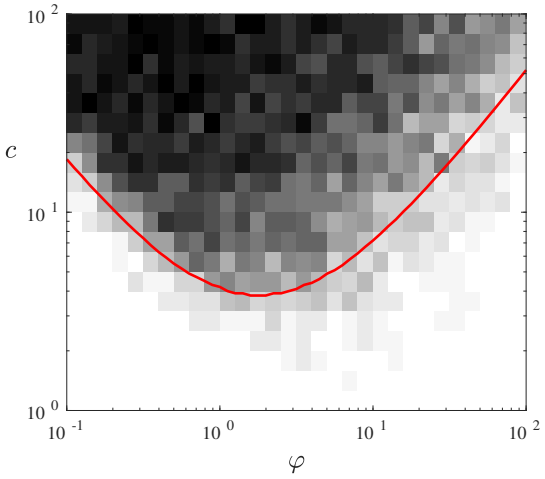


Figure 3: The density plot shows the probability (estimated as a fraction of 25 simulations per pixel) of an outbreak of size > 100, starting from a single infected node, in an ER network of 10000 nodes. The red line is the boundary of the region where survival of the coupled branching process is possible.

In this talk I will introduce a multi-type branching process which couples to the secondary infection — an approach that has previously enjoyed success in approximating SIR-type models in large populations as seen for instance in [2, 9], the effectiveness here demonstrated in figure 3. I will then describe how study of this process enables characterisation of the phase diagram in figure 2, concluding with some details of the proof of theorem 1.

References

[1] WHO — Hepatitis D. *WHO*, 2017.

[2] R. Bartoszynski. Branching processes and the theory of epidemics. In *Proceedings of the Fifth Berkeley symposium on Mathematical Statistics and Probability*, pages 259–69, 1967.

[3] R. A. Blythe, R. J. Allen, et al. Parasites on parasites: Coupled fluctuations in stacked contact processes. *Europhysics Letters*, 101(5):50001, 2013.

[4] M. M. Henkel, H. Hinrichsen, S. Lubeck, and M. Pleimling. *Non-equilibrium phase transitions.* Springer, 2008.

[5] N. Lanchier and Y. Zhang. Some rigorous results for the stacked contact process. *arXiv preprint arXiv:1410.3842v1*, 2014.

[6] M. E. Newman. Spread of epidemic disease on networks. *Physical review E*, 66(1):016128, 2002.

[7] M. E. Newman and C. R. Ferrario. Interacting epidemics and coinfection on contact networks. *PloS one*, 8(8):e71321, 2013.

[8] T. Rogers. Assessing node risk and vulnerability in epidemics on networks. *Europhysics Letters*, 109(2):28005, 2015.

[9] S. Singh. *Branching processes in disease epidemics*. Cornell University, 2014.

EQUILIBRIUM COMMUNITY STRUCTURE IN BINARY-STATE ADAPTIVE VOTER MODELS

Philip S. Chodrow, Peter J. Mucha

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Adaptive Voter Models (AVMs) provide a natural framework for studying the coevolution of node states and network topology. We develop a novel analytic framework for these systems, which outperforms existing methods in both accuracy and computational tractability. The strong scaling properties of this framework open the door to analyses of previously-intractable systems with large mean degrees and multiple, possibly structured node states.

Model Description

Community structure is a fundamental object of study in a broad range of behavioral and biological networks, but canonical models are largely silent on the processes by which communities form and persist as networks evolve in time. Seeking a dynamical story of community structure, we study an adaptive voter model introduced by [6] and further studied by [4] and [3] that couples node opinions and network topology. In the classical model, agents update 0-1 opinions in response to those of their neighbors ("voting") or cut ties to a disagreeing neighbor in favor of a new one selected uniformly at random ("rewiring"). These systems display an interesting and frequently-studied *fragmentation transition*: for sufficiently high values of the rewiring rate α the system splits into two equally-sized connected components of 0s and 1s, while for lower values of α the voting process breaks symmetry between the component sizes. As models of nontrivial community structure, however, the classical AVM is unsatisfactory, as its equilibrium states are fully fragmented with no edges between disagreeing nodes.

We therefore study a noisy AVM [5] in which nodes are subject to external noise that randomly mutates their opinions at rate λ. This system is ergodic, and we may therefore study its equilibrium distribution, a sample from which is illustrated in Figure 1. As a simple measure of community structure, we consider the expected modularity q at equilibrium. In a rewire-to-random model, q is a solution to

$$0 = 2\lambda c q - (1-\lambda)\frac{\alpha}{2} + (1-\lambda)(1-\alpha)V \,, \qquad (1)$$

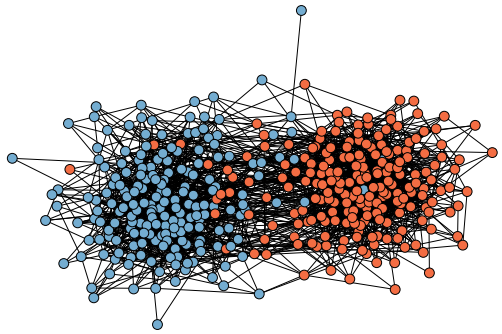


Figure 1: Sample at equilibrium from a rewire-to-random AVM with $N = 400$ agents, mean degree $c = 10$, rewiring rate $\alpha = 0.9$, and mutation rate $\lambda = 0.01$. The modularity of this network is $q = 0.38$, compared to a maximum possible value of $q_{\max} = 0.5$. Nodes are colored according to their opinion and drawn under a spring layout.

where V is the expected change in modularity due to a voting event, and is a function of the entire graph state. Approximating this term is the fundamental analytical challenge in studying AVMs.

Methods and Results

We estimate V in the fragmented regime ($q = \frac{1}{2}$) as follows. We consider a node that has just changed its opinion, generating a set of discordant edges. We track each of these edges until they are resolved, logging voting events along the way. The mean change in discordant edges per voting event, labeled $\hat{V}(q)$, is our estimate for V. The Markovian approximation consists in supposing that $\hat{V}$ is a function of q and the system parameters – no higher moments are used.

The simplified structure of the scenario enables us to compute $\hat{V}\left(\frac{1}{2}\right)$ analytically in terms of α, λ, and the mean degree c. Numerically solving (1) when $q = \frac{1}{2}$ and replacing V with $\hat{V}\left(\frac{1}{2}\right)$ yields a prediction for α^*, the critical rewiring rate past which $q < \frac{1}{2}$. Figure 2 compares

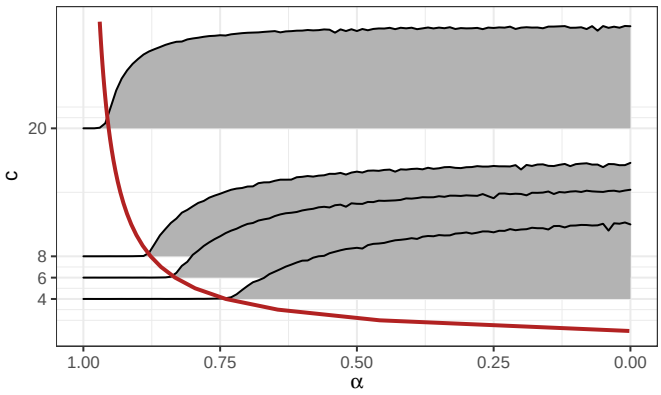


Figure 2: Phase transition in $\rho = \frac{1}{2} - q$ as a function of the rewiring rate α. Grey densities give the observed ρ sampled at equilibrium from AVMs with $\lambda = 2^{-10}$, $N = 10^5$, and varying mean degree c. The red curve gives the theoretical prediction.

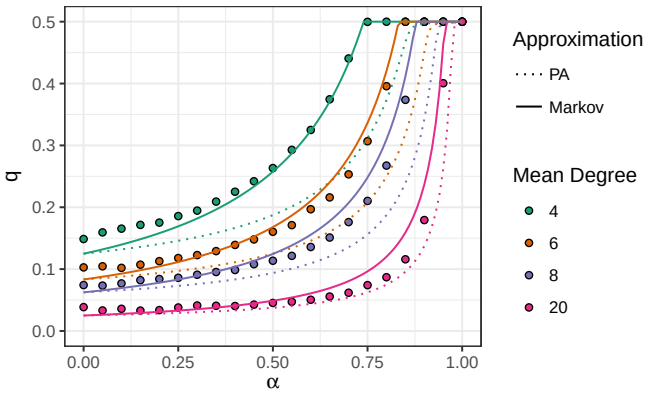


Figure 3: Points give the observed modularity q as a function of the rewiring rate α for varying mean degree for a system with $\lambda = 2^{-10}$, $N = 10^5$, and varying mean degree c.

this prediction to data for a wide range of c in the case $\lambda = 2^{-10}$, finding the agreement to be nearly exact.

To estimate the modularity past the phase transition in α, we employ a simple linear interpolation:

$$\hat{V}(q) \approx \frac{\hat{V}(q_0) - \hat{V}\left(\frac{1}{2}\right)}{2q_0 - 1}(2q - 1) + \hat{V}\left(\frac{1}{2}\right) , \qquad (2)$$

where q_0 is the steady-state modularity when $\alpha = 0$. Both q_0 and $\hat{V}(q_0)$ may be estimated via generating functions [1]. The resulting approximation for q is shown in Figure 3.

As compared to the common Pair Approximation (dashed lines), equation 2 much more accurately estimates both the phase transition α^* and the expected modularity past α^*. Compared to a more sophisticated approach such as approximate master equations [4], ours is both more accurate and much faster to compute, requiring the solution of a single nonlinear equation rather than $\Theta(c^2)$ equations. Similarly, our approach displays comparable accuracy to the active motif approximation of [3] for the rewire-to-same model variant, but does not require the numerical location of zeros in the spectrum of a $c^2 \times c^2$ matrix.

Discussion

The simplicity of our analytical approach allows the consideration of systems of greater complexity than previously possible. The most promising generalization is to multiple opinion states; while considered in [2, 7], to our knowledge no analytic methods exist for modeling supercritical behavior in these systems. Another promising generalization is to structured node state spaces, such as opinions lying on a spectrum from "very conservative" to "very liberal." We are optimistic that the methods developed here may eventually assist in the analyses of such models.

References

[1] B. Allen, A. Traulsen, C. E. Tarnita, and M. A. Nowak. How mutation affects evolutionary games on graphs. *Journal of Theoretical Biology*, 299:97–105, 2012.

[2] G. A. Böhme and T. Gross. Fragmentation transitions in multi-state voter models. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(6), 2012.

[3] G. Demirel, F. Vazquez, G. A. Böhme, and T. Gross. Moment-closure approximations for discrete adaptive networks. *Physica D: Nonlinear Phenomena*, 267:68–80, 2014.

[4] R. Durrett, J. P. Gleeson, A. L. Lloyd, P. J. Mucha, F. Shi, D. Sivakoff, J. E. S. Socolar, and C. Varghese. Graph fission in an evolving voter model. *Proceedings of the National Academy of Sciences of the United States of America*, 109(10):3682–7, 2012.

[5] M. Ji, C. Xu, C. W. Choi, and P. M. Hui. Correlations and analytical approaches to co-evolving voter models. *New Journal of Physics*, 15, 2013.

[6] D. Kimura and Y. Hayakawa. Coevolutionary networks with homophily and heterophily. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 78(1):1–7, 2008.

[7] F. Shi, P. J. Mucha, and R. Durrett. Multiopinion coevolving voter model with infinitely many phase transitions. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 88(6):1–15, 2013.

CONJOINING UNCOOPERATIVE SOCIETIES TO PROMOTE EVOLUTION OF COOPERATION

Babak Fotouhi[1], *Naghmeh Momeni*[1,2], *Benjamin Allen*[1,3], *Martin A. Nowak*[1]

[1] *Program for Evolutionary Dynamics, Harvard University* [2] *MIT Sloan School of Management* [3] *Department of Mathematics, Emmanuel College*

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Network structure affects the evolution of cooperation. We study analytically how cooperation-inhibiting networks can be conjoined to form composite cooperation-promoting structures. To this end, we introduce a newly-discovered exact method based on the equivalence between evolutionary games on graphs and coalescing random walks [1].

Introduction

In diverse social and organizational contexts, there are nodes who span the boundaries of otherwise-segregated groups, and bridge the 'structural holes' of the networks via long ties. By sewing the patches of the network together, these nodes facilitate cooperation and collective functioning [2–4]. In our main paper, we explore how such conjoinings of various cooperation-hindering topologies—random, non-random, and real—can rescue cooperation. In this abstract, we aesthetically selected a few non-random cases to present.

Model

We use the framework of evolutionary graph theory [1]. Nodes play the sequential 'donation game' version of Prisoner's Dilemma (PD) on a network (results are generalizable to arbitrary games). The payoff matrix is $b-1, -1, b, 0$. We consider the voter-type updating dynamics (Death-Birth) in the limit of weak selection. We calculate the critical benefit-to-cost ratio b^* such that for $b > b^*$ natural selection favors cooperation over defection.

We consider simple graphs—connected, unweighted, undirected, loop-less graphs without multiple edges. Denote the number of nodes with N, the degree of node x with k_x, and its set of neighbors by $\mathcal{N}_x$. Then, we define: $p_x := (1/k_x)\sum_{y\in\mathcal{N}_x}(1/k_y)$. We then solve the following system of linear equations for coalescence times τ_{xy}, which are the expected remeeting times of two random walkers starting form nodes x and y:

$$\tau_{xy} = \tau_{yx} = (1-\delta_{xy})\left[1 + \frac{1}{2k_x}\sum_{z\in\mathcal{N}_x}\tau_{zy} + \frac{1}{2k_y}\sum_{z\in\mathcal{N}_y}\tau_{zx}\right]$$

Then, we have: $b^* = \dfrac{\sum_x\sum_{y\in\mathcal{N}_x}(\tau_{xy}-k_x)}{\sum_x\sum_{y\in\mathcal{N}_x}(\tau_{xy}p_{xy}+1-2k_x)}$

Selected Results

For space limitations, we present the simplified version of the results: limiting behaviors in the large-size limits. We only present results for cliques, stars, and rich clubs. For **cliques** (complete graphs), $b^* < 0$. Thus natural selection favors defection regardless of b—these structures promote *spite*. If we conjoin two cliques via a 'chain junction' (Fig. 1a), then natural selection favors cooperation in the composite graph. With direct connection, or with one intermediary node, b^* grows as n^2. Remarkably, with two (or more) intermediary nodes, the leading behavior of b^* drops to linear. For the 'star junction' (Fig. 1b) also b^* grows linearly with n. For more than two cliques, we present the ring structure (Fig. 1c), star structure (Fig. 1d), and hierarchical organization (Fig. 1e). For the **star** graph, b^* in infinite. Connecting stars in the above-said manners lead to structures whose limiting behavior is *constant*, i.e., independent of network size. These structures are strong promoters of cooperation. Here we only present direct interconnection (Fig. 1f). Finally, we present **rich clubs**, which consist of a clique (core) and a periphery: peripheral nodes are not connected together but each of them is connected to every core node. For rich clubs, $b^* < 0$, so it promotes spite. Conjoining rich clubs gives structures whose b^* approaches a constant in the limit of large periphery. A simple example is depicted in Fig. 1g.

References

[1] B. Allen, G. Lippner, Y.-T. Chen, B. Fotouhi, N. Momeni, S.-T. Yau, and M. A. Nowak. Evolutionary dynamics on any population structure. *Nature*, 544(7649):227–230, 2017.

[2] R. S. Burt. *Structural holes: The social structure of competition.* Harvard university press, 2009.

[3] F. Fukuyama. Social capital, civil society and development. *Third world quarterly*, 22(1):7–20, 2001.

[4] R. D. Putnam. *Bowling alone: The collapse and revival of American community.* Simon and Schuster, 2001.

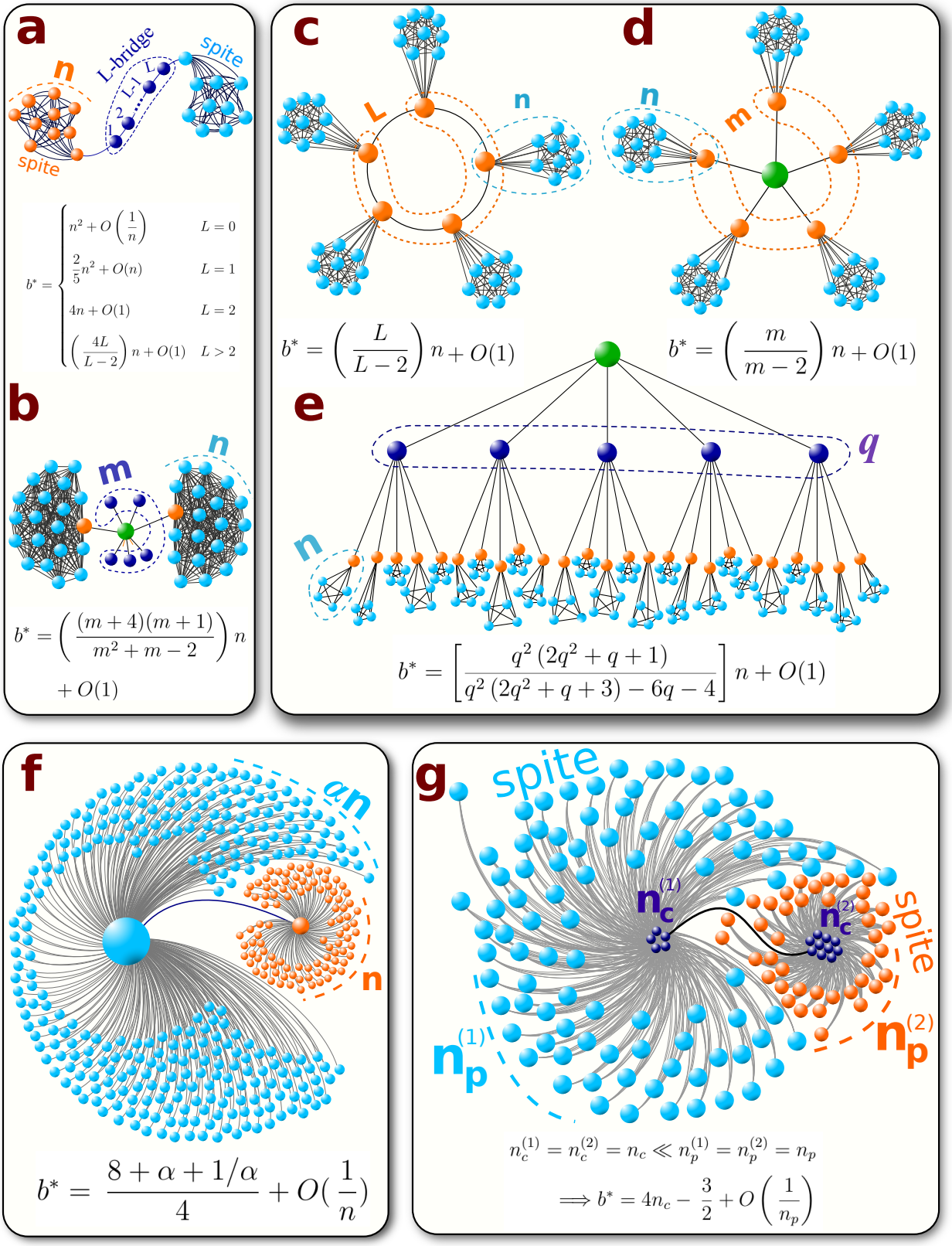


Figure 1: **Conjoining of different topologies.** (a) two cliques connected via a 'chain junction', (b) two cliques conjoined via a 'star unction', (c) multiple cliques connected on a ring structure, (d) multiple cliques connected in a star structure, (e) multiple cliques organized on a hierarchical structure, (f) conjoining two stars, (g) conjoining two rich-club graphs via the cores.

SELF-ORGANIZATION OF DRAGON KINGS

Yuansheng Lin, Keith Burghardt, Martin Rohden, Pierre-André Noël, Raissa M. D'Souza

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

The mechanisms underlying cascading failures are often modeled via the paradigm of self-organized criticality. Here we introduce a one-parameter model that leads to "Dragon Kings", which are massive failures caused by mechanisms distinct from smaller failures, throughout the parameter regime. We demonstrate that the size of the initial failed weak cluster predicts the likelihood of a Dragon King event with high accuracy and we develop a simple control strategy which reduces the number of Dragon Kings by orders of magnitude.

Abstract

Many natural and engineered systems exhibit rare, catastrophic events [19, 20, 5, 1, 2, 6, 12, 23, 14, 22, 9]. Two categories for such events have been proposed: Black Swans, which are tail events in a power-law distribution, and Dragon Kings (DKs), which are outliers involving mechanisms absent in smaller events that occur far more frequently than a power-law would predict. The power-law distribution necessary for Black Swans to exist is often explained by self-organized criticality (SOC): a tug-of-war that poises the system close to a critical point without any need for tuning of external parameters [2, 3, 6, 12]. Although the prediction of Black Swans can sometimes beat random chance [17], the task appears inherently difficult [11, 21]. Despite this drawback, there are simple methods to push SOC systems away from criticality, thus reducing the size of Black Swans [4, 15, 12].

It has been proposed that DKs occur in systems that have low heterogeneity and strong coupling (as defined in [16]) and that, in contrast, Black Swans occur in systems with weaker coupling and higher heterogeneity. Whereas Black Swans often have no associated length- and time-scales, DK events do: there are typical places and times when DKs will and will not occur. This has been successfully applied to, for example, prediction of material failure and crashes of stock markets [19], and has been seen in engineered systems, such as error cascades in a collection of robots [10]. Unlike Black Swans, however, it has been an open problem to control DKs in many situations and to elucidate the mechanisms underlying these often self-amplifying cascades [20]. Recent advances on controlling DKs have been based on low-dimensional models, such as coupled oscillators [7], but control of DKs in models of high-dimensional complex systems has been lacking.

We introduce a simple model, which we call the "*complex contagion*" (CC) model [8], where nodes in a network self-organize to be "weak" or "strong" to failure, capturing the tradeoffs between degradation and reinforcement of elements in a system. The model has properties in common with many complex engineered systems, such as power grids [18] and fault trees [13]. In our model the initial failure of a random weak node can lead to a cascade of subsequent node failures. A weak node fails as soon as *one* of its neighbors fails, and a failed weak node has a small probability, ϵ, to be reinforced and upgraded to a strong node upon repair. Strong nodes fail as soon as *two* of their neighbors fail, which is similar to complex contagions, thus motivating the name for our model. Strong nodes also independently degrade (i.e., become weak) at a slow rate. This model can lead to self-amplifying failures that cascade across clusters of weak nodes. The CC model is to our knowledge the simplest model that produces self-amplifying cascading failures.

Each cascade causes small changes in the number of weak and strong nodes, which drives the model to a state that spontaneously generates DKs (failures of nearly the entire system) over all values of $\epsilon < 1$, a prediction that we have confirmed for ϵ over several orders of magnitude. The DKs created in the CC model are due to "cluster hopping" cascades, in which failures hop across multiple weak-node clusters. These large cascades, however, are only likely to occur when the first cluster of weak nodes to fail is larger than some threshold. We use logistic regression to fit the probability of DKs in simulations given the size of the initial weak cluster, and discover that we can predict large DKs with startling accuracy: the area under the receiver operating characteristic is always close to the maximum value of 1 across several values of

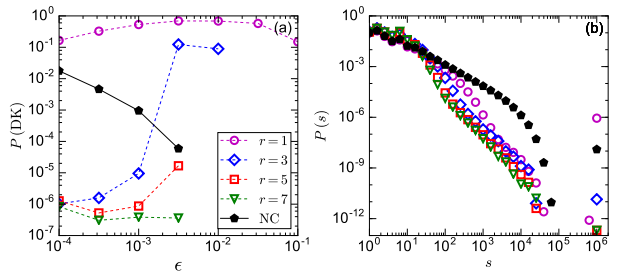


Figure 1: (Color online) Controlling DKs. (a) The probability a DK occurs per timestep versus ϵ in both the non-controlled scenario (NC, pentagons), and in the controlled scenario with r weak nodes chosen: $r = 1$ (circles), $r = 3$ (diamonds), $r = 5$ (squares) and $r = 7$ (triangles). Simulations are realized for $N = 10^6$, and standard errors are smaller than marker sizes. See main text for details of the control method. (b) Failure size distributions in both the non-controlled scenario and in the controlled scenario for $\epsilon = 3.2 \times 10^{-4}$.

ϵ. We take advantage of this finding to control whether a small initial failure will cascade into a DK event, see Fig. 1. In this figure, we simulated the model for $10N$ timesteps, and then added the control condition: at each timestep, we pick r weak nodes, and reinforce the weak node connected to the largest weak-node cluster. This control condition can decrease DKs and large cascades by orders of magnitude, even though the number of weak and strong nodes remains the same on average. Interestingly, reinforcing a weak-node at random ($r = 1$) can instead make DKs more likely than the non-controlled condition.

In conclusion, we find a simple model that creates DKs over a wide parameter space, and find a simple method to predict and control these DKs based on how they develop from small initial cascades.

References

[1] P. Bak. *How Nature Works: the Science of Self-organized Criticality*. Copernicus, New York, 1996.

[2] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality: An explanation of 1/f noise. *Phys. Rev. Lett.*, 59(4):381–384, 1987.

[3] P. Bak, C. Tang, and K. Wiesenfeld. Self-organized criticality. *Phys. Rev. A*, 38(1):364–374, 1988.

[4] D. O. Cajueiro and R. F. S. Andrade. Controlling self-organized criticality in sandpile models. *Physical Review E*, 81(1):015102, 2010.

[5] J. M. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. *Phys. Rev. E*, 60(2):1412–1427, 1999.

[6] B. A. Carreras, V. E. Lynch, I. Dobson, and D. E. Newman. Critical points and transitions in an electric power transmission model for cascading failure blackouts. *Chaos: An interdisciplinary journal of nonlinear science*, 12(4):985–994, 2002.

[7] H. L. d. S. Cavalcante, M. Oriá, D. Sornette, E. Ott, and D. J. Gauthier. Predictability and suppression of extreme events in a chaotic system. *Physical review letters*, 111(19):198701, 2013.

[8] D. Centola and M. Macy. Complex contagions and the weakness of long ties. *American journal of Sociology*, 113(3):702–734, 2007.

[9] R. M. D'Souza. Curtailing cascading failures. *Science*, 358(6365):860–861, 2017.

[10] M. Gauci, M. E. Ortiz, M. Rubenstein, and R. Nagpal. Error cascades in collective behavior: A case study of the gradient algorithm on 1000 physical agents. In *Proceedings of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017)*, pages 1404–1412. International Foundation for Autonomous Agents and Multiagent Systems, 2017.

[11] R. J. Geller, D. D. Jackson, Y. Y. Kagan, and F. Mulargia. Earthquakes cannot be predicted. *Science*, 275(5306):1616, 1997.

[12] H. Hoffmann and D. W. Payton. Suppressing cascades in a self-organized-critical model with non-contiguous spread of failures. *Chaos Soliton. Fract.*, 67:87 – 93, 2014.

[13] W. S. Lee, D. L. Grosh, F. A. Tillman, and C. H. Lie. Fault tree analysis, methods, and applications – a review. *IEEE TRANSACTIONS ON RELIABILITY*, 34:194–203, 1985.

[14] J. Lorenz, S. Battiston, and F. Schweitzer. Systemic risk in a unifying framework for cascading processes on networks. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):441–460, 2009.

[15] P.-A. Noël, C. D. Brummitt, and R. M. D'Souza. Controlling self-organizing dynamics on networks using models that self-organize. *Phys. Rev. Lett.*, 111:078701, 2013.

[16] I. Osorio, M. G. Frei, D. Sornette, J. Milton, and Y.-C. Lai. Epileptic seizures: quakes of the brain? *Physical Review E*, 82(2):021919, 2010.

[17] O. Ramos, E. Altshuler, and K. J. Måløy. Avalanche prediction in a self-organized pile of beads. *Physical review letters*, 102(7):078701, 2009.

[18] T. A. Short. *Electric Power Distribution Handbook*. CRC Press, Boca Raton, Fl, 2004.

[19] D. Sornette. Dragon-kings, black swans and the prediction of crises. *International Journal of Terraspace Science and Engineering*, 2(1):1–18, 2009.

[20] D. Sornette and G. Ouillon. Dragon-kings: mechanisms, statistical methods and empirical evidence. *Eur. Phys. J. Special Topics*, 205:53–64, 2012.

[21] N. Taleb. *The Black Swan: The Impact of the Highly Improbable*. Random House, 2007.

[22] C. J. Tessone, A. Garas, B. Guerra, and F. Schweitzer. How big is too big? critical shocks for systemic failure cascades. *Journal of Statistical Physics*, 151(3-4):765–783, 2013.

[23] S. Wheatley, B. Sovacool, and D. Sornette. Of disasters and dragon kings: a statistical analysis of nuclear power incidents and accidents. *Risk analysis*, 37(1):99–115, 2017.

RANDOM SPATIAL NETWORKS AND ANALYTIC MODELS FOR THE SPREAD OF SIR EPIDEMICS

Joel C. Miller

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We describe a class of random spatial networks which allow us to control the edge-length distribution and the degree distribution of the nodes. Because of their structure, these networks permit analytic investigation. We derive an analytic model for SIR disease spread on these networks, and we use the model to predict properties such as the speed of traveling waves.

This material is based primarily on [8].

Introduction

In many important networks, for example human contacts [7], neurons in the brain [2, 10], wireless sensor networks [4], protected plant/animal habitats [6], wildlife interaction networks [3, 5], and even the physical internet [12], the connections are not randomly distributed in space. Rather, shorter connections are favoured.

Despite the frequency of spatial networks, there has generally been a lack of network models which allow us to simultaneously control the degree distribution and the distribution of partnership distances, while still allowing us to develop analytic models for processes spreading on those networks.

The Random Spatial Network class

It is relatively simple to build a network from our Random Spatial Network class. To do so, we

- place nodes into a space V with density ρ,
- assign each node u an expected degree κ_u from some distribution of expected degrees $P(\kappa)$,
- and then join any two nodes u and v at distance d_{uv} with probability

$$p_{uv} = \min\left(1, \frac{\kappa_u \kappa_v f(d_{uv})}{\rho \langle K \rangle}\right)$$

where $\langle K \rangle$ is the average of κ and $f(d_{uv})$ is a distance kernel measuring the impact of distance on partnership formation (the integral of f over V is 1).

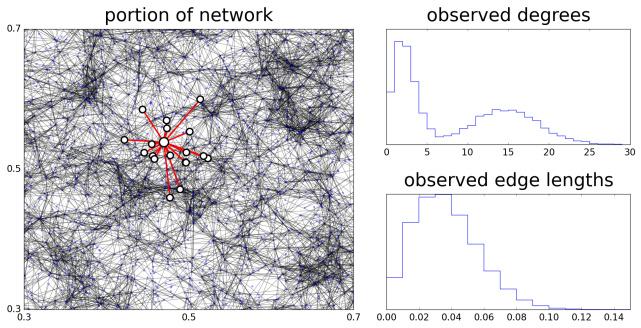


Figure 1: An example RSN and its properties. The distance kernel is a Gaussian, $f(d) = \exp(-d^2/2\sigma^2)/2\pi\sigma^2$ with $\sigma = 0.03$. The imposed distribution of expected degrees is $P(2) = P(15) = 0.5$. The density is $\rho = 10000$. One node and its neighbors are highlighted. A random network without spatial structure would exhibit neighbors throughout the domain.

This is a special case of the inhomogeneous random graph class [1]. Fig. 1 shows a sample RSN and some properties.

Note that at high density, two neighbors v and w of a given node u are unlikely to be neighbors of one another because p_{vw} scales like $1/\rho$. Thus the clustering coefficient scales like $1/\rho$.

By playing with the density and distance kernel, we can create a wide range of networks. An interesting case is networks that have many properties of small-world networks, but have a negligible level of clustering. These are formed by having a high density and the vast majority of edges have short distance, but a few are long. Many dynamic processes on these networks can display effects typically considered to be small-world.

Analytic SIR model

It is possible to use "edge-based compartmental modeling" [9] to derive an analytic model for the spread of Susceptible–Inected–Recovered (SIR) disease (if stochastic

fluctuations are negligible). We find [8]

$$\frac{\partial}{\partial t}\Theta(\vec{x},t) = -\beta\Theta(\vec{x},t) + \gamma(1-\Theta(\vec{x},t)) + \beta\frac{\int_V S(\hat{\vec{x}},0)\Psi'(\Theta(\hat{\vec{x}},t))f(|\hat{\vec{x}}-\vec{x}|)\,\mathrm{d}\hat{\vec{x}}}{\langle\kappa\rangle}$$

$$S(\vec{x},t) = S(\vec{x},0)\Psi(\Theta(\vec{x},t)),$$

$$\frac{\partial}{\partial t}R(\vec{x},t) = \gamma(1-S(\vec{x},t)-R(\vec{x},t))$$

$$I(\vec{x},t) = 1-S(\vec{x},t)-R(\vec{x},t)$$

where

$$\Psi(\Theta(\vec{x},t)) \equiv \int_0^\infty e^{-\kappa(1-\Theta(\vec{x},t))}P(\kappa)\,\mathrm{d}\kappa.$$

Comparison of simulation and theory are good in the high-density limit, as seen in Fig. 2.

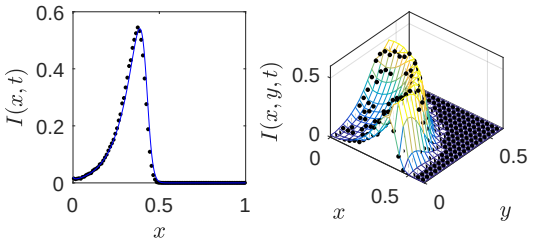


Figure 2: Comparison of stochastic simulation of SIR disease dynamics with numerical solution of analytic equations on 1D and 2D random spatial networks.

The analytic model leads to an explicit calculation for the wavespeed. So long as f decays at least exponentially fast at large distance, the predicted wavespeed is finite, though Fig. 3 shows that convergence to this wave speed is quite slow as node density increases. If the tail decays slower than exponentially, then the disease spread is dominated by "hop-and-spread" dynamics.

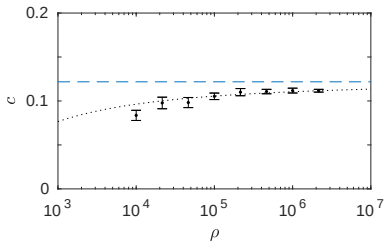


Figure 3: Comparison of wave speed for SIR disease dynamics observed in stochastic simulations (small black circles) and analytic prediction (dashed cyan line). For the stochastic simulations, we generate $n_{rep} = 25$ 1D spatial networks for each node density ρ using a distance kernel $f(|x|) = \phi_{0,0.01}(|x|)$, where $\phi_{\mu,\sigma}(x)$ is the probability density function for the normal random variable with mean μ and standard deviation σ. All nodes have expected degree $\kappa = 10$. For each network we realize one SIR simulation with disease parameters $\beta = 1$ and $\gamma = 3$. The black circles show the average wave speed resulting from 25 network realizations (vertical bars indicate 95% confidence intervals). The dotted black curve represents the expected convergence behavior of $c^* - K/\ln^2\rho$ [11] where the constant K is obtained by fitting the curve to the rightmost black circle.

References

[1] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3, 2007.

[2] E. Bullmore and O. Sporns. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, 10(3):186–198, 2009.

[3] S. Davis, B. Abbasi, and *et al.* Spatial analyses of wildlife contact networks. *J. of The Royal Society Interface*, 12(102):20141004, 2015.

[4] M. Haenggi, J. G. Andrews, and *et al.* Stochastic geometry and random graphs for the analysis and design of wireless networks. *Selected Areas in Communications, IEEE Journal on*, 27(7):1029–1046, 2009.

[5] R. K. Hamede, J. Bashford, H. McCallum, and M. Jones. Contact networks in a wild tasmanian devil (Sarcophilus harrisii) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecology Letters*, 12(11):1147–1157, 2009.

[6] I. Hanski and O. Ovaskainen. The metapopulation capacity of a fragmented landscape. *Nature*, 404(6779):755–758, 2000.

[7] R. Lambiotte, V. D. Blondel, and *et al.* Geographical dispersal of mobile communication networks. *Physica A*, 387(21):5317–5325, 2008.

[8] J. C. Lang, J. L. Kaiser, H. D. Sterck, and J. C. Miller. Analytic models for sir disease spread on random spatial networks. *Journal of Complex Networks*, 2018.

[9] J. C. Miller, A. C. Slim, and E. M. Volz. Edge-based compartmental modelling for infectious disease spread. *Journal of the Royal Society Interface*, 9(70):890–906, 2012.

[10] R. O'Dea, J. J. Crofts, and M. Kaiser. Spreading dynamics on spatially constrained complex brain networks. *J. of The Royal Society Interface*, 10(81):20130016, 2013.

[11] D. Panja. Effects of fluctuations on propagating fronts. *Physics Reports*, 393(2):87–174, 2004.

[12] S.-H. Yook, H. Jeong, and A.-L. Barabási. Modeling the internet's large-scale topology. *Proceedings of the National Academy of Sciences*, 99(21):13382–13386, 2002.

NETWORK CONSTRAINTS ON LEARNABILITY OF PROBABILISTIC MOTOR SEQUENCES

Ari E. Kahn, Elisabeth A. Karuza, Jean M. Vettel, and Danielle S. Bassett

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Extensive evidence suggests that learners are sensitive to first order statistics embedded in sequential input; however, the impact of higher-order structure remains unclear, especially in the context of a motor learning task. We addressed this by formalizing temporal statistical learning using a graph-based framework. We found both that learners were sensitive to meso-scale structure, and that certain node-level features were associated with varying levels of learning difficulty.

Introduction

Our ability to interact with our environment necessitates that we parse complex stimuli into smaller units, such as words and phrases in language input, or events in streams of visual stimuli. This essential process relies at least in part on the statistical regularities present around us, and often operates automatically and without any explicit, verbalizable knowledge of underlying rules. As early as infancy, humans reliably detect the probabilities with which one stimulus transitions to another (transition probabilities, such as one syllable following another in speech) and the frequencies with which stimuli co-occur[7]. Similar forms of pattern sensitivity have been observed beyond the language domain, including motor learning[6] and visual event segmentation[10].

Learning Based on Higher-Order Statistics

Evidence suggests that, depending on context, learners can extract both adjacent and non-adjacent dependencies between stimuli[2]. Ongoing research continues to examine which types of statistics facilitate learning, as well as the effect of presenting these statistics in various domains (e.g., in the context of complex associations between button presses [1] or social stimuli[9]). Together, these studies suggest that second- or third-order statistical relationships may be encoded implicitly, and furthermore, that higher-level organizational principles themselves might be implicitly learned. Indeed, recent work has shown that temporal ordering of visual stimuli can convey the organizational principle of modularity[8]. This observation opens up the possibility of studying whether certain broad-scale organizational structures might best facilitate learning[4].

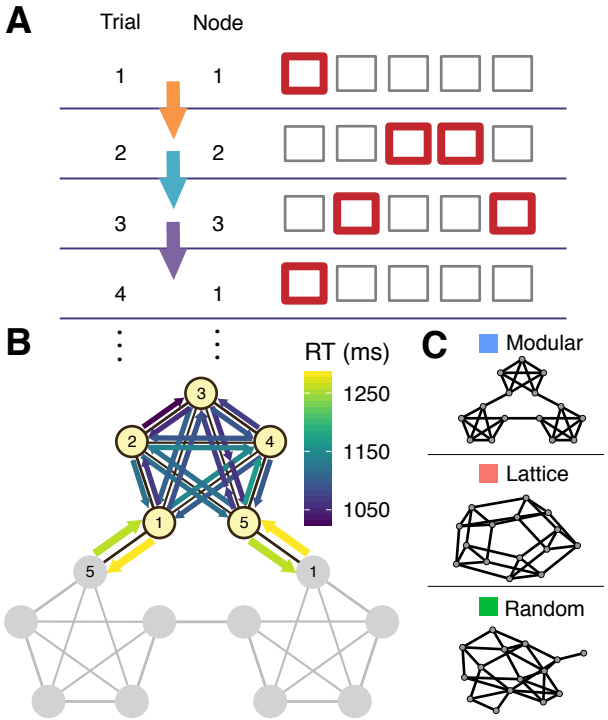


Figure 1: Experimental Design. *(A)* An example of the first few trials of a graph traversal. Each node is uniquely assigned a key combination (red squares). The sequence of key combinations is determined by a walk on the graph. *(B)* Example modular graph composed of three clusters. Colors correspond to the mean reaction time (RT) when traversing the indicated edge. The increase in RT between clusters is apparent, here visualized by yellower colors on the edges that connect the top cluster with the two on bottom. *(C)* The three graph structures that we examined: a modular graph, a lattice graph, and a random graph.

Formalization with Network Science

An ideally suited language in which to define such higher-order principles is network science[5], an emerging interdisciplinary field that addresses the architecture, dynamics,

and design of complex systems composed of many connected parts. In the context of learning, we can construct a graph that encodes the pattern of relationships between objects, movements, or sounds. We recently capitalized on this approach to define a graph from which the temporal ordering of visual stimuli was drawn[3]. Learners exhibited a strong sensitivity to graph structure, as indexed by an increase in reaction time when transitioning between modules (i.e., a suprisal effect). Importantly, this effect was dependent on the type of traversal through the graph, and was more strongly pronounced when traversals through the graph provided redundancy in local information.

Experimental Design

Here, we asked whether the higher-order architecture of three distinct graph structures influenced learning of temporal sequences. Specifically, we trained subjects on a probabilistic motor task, where each trial was drawn from a traversal through a graph. Each node represented a stimulus, and each edge represented a possible transition between two nodes (Fig. 1). We contrasted walks on a clusted graph with modular organization, an unclustered lattice graph, and a random graph without any regular structure.

Sensitivity to Meso- and Micro-scale Structure

Based on prior work[3], we hypothesized that learners would gain an implicit representation of the modular graph structure as evidenced by a surprisal effect (Fig. 1B). Next, we systematically varied graph structure to examine the impact of graph topology on the acquisition of complex, multi-element motor sequences (Fig. 2A). Results indicate that learning, indexed here by participants' overall reaction times, was strongly mediated by the graph's meso-scale organization, with modular graphs being associated with shorter reaction times than random and lattice graphs. Interestingly, variations in a node's number of connections (degree) and a node's role in mediating long-distance communication (betweenness centrality) impacted graph learning, even after accounting for level of practice on that node (Fig. 2B). These results demonstrate that the graph architecture underlying temporal sequences of stimuli fundamentally constrains learning, and network science provides a valuable framework for assessing how learners encode complex, temporally structured information.

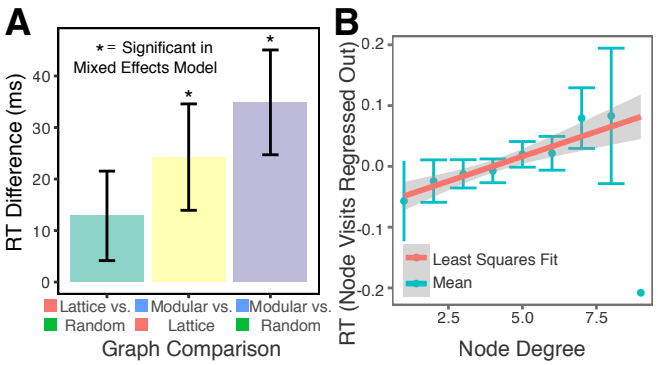


Figure 2: Effects of Graph Structure. *(A)* Differences in RT by graph type, across graphs learned in sequence, showing milliseconds by which the modeled effect for the top listed graph is faster. The increase in RT from lattice to modular and from random to modular graphs are both significant to $p = 0.02$ and $p = 0.001$, respectively. Examples of the graph types are shown to the right. *(B)* Mean RT shown as a function of degree, where the mean was z-scored across the 15 nodes for a given subject.

References

[1] A. Cleeremans and J. L. McClelland. Learning the structure of event sequences. *Journal of Experimental Psychology: General*, 120(3):235–253, 1991.

[2] R. L. Gómez. Variability and Detection of Invariant Structure. *Psychological Science*, 13(5):431–436, sep 2002.

[3] E. A. Karuza, A. E. Kahn, S. L. Thompson-Schill, and D. S. Bassett. Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, 7(1):1–9, 2017.

[4] E. A. Karuza, S. L. Thompson-Schill, and D. S. Bassett. Local Patterns to Global Architectures: Influences of Network Topology on Human Learning, 2016.

[5] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[6] M. J. Nissen and P. Bullemer. Attentional requirements of learning: Evidence from performance measures. *Cognitive Psychology*, 19(1):1–32, jan 1987.

[7] J. R. Saffran, R. N. Aslin, and E. L. Newport. Statistical Learning by 8-Month-Old Infants. *Science*, 274(5294):1926–1928, 1996.

[8] A. C. Schapiro, T. T. Rogers, N. I. Cordova, N. B. Turk-Browne, and M. M. Botvinick. Neural representations of events arise from temporal community structure. *Nature Neuroscience*, 16(4):486–492, feb 2013.

[9] S. Tompson, A. Kahn, E. Falk, J. Vettel, and D. Bassett. Individual Differences in Learning Social and Non-Social Network Structures. *arXiv*, pages 1–42, 2017.

[10] N. B. Turk-Browne, J. A. Jungé, and B. J. Scholl. The Automaticity of Visual Statistical Learning. *Journal of Experimental Psychology: General*, 134(4):552–564, 2005.

MAXIMUM ENTROPY SPARSE RANDOM GRAPHS WITH GIVEN AVERAGE DEGREE AND CLUSTERING

Pim van der Hoorn, Johan van Leeuwaarden, Gabor Lippner and Dmitri Krioukov

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Statistical analysis of real-world networks requires random graph models that generated graphs with structural constraints, corresponding to the network of interest, while being maximally unbiased with respect to all other structural properties. In particular, we are interested in such models that have scale-free degree distribution and strong clustering, since these structures are common among most complex networks. In this work we take a first important step towards this goal by analyzing the structure of sparse maximum entropy graphs with given average degree and clustering. We find that the solution is given by a symmetric sparse stochastic block model.

Maximum entropy problem for graphs

For any average degree value k and average number of triangles per node t such that $k > \sqrt{2t}$ we want to characterize the sequence $(\mathcal{G}_n)_{n\geq 1}$ of maximally unbiased random graphs that satisfy the following limit constraints:

$$\lim_{n\to\infty} \mathbb{E}\left[D(\mathcal{G}_n)\right] = k \qquad \lim_{n\to\infty} \mathbb{E}\left[T(\mathcal{G}_n)\right] = t, \tag{1}$$

where $D(\mathcal{G}_n)$ denotes the degree of a node sampled uniformly at random in $\mathcal{G}_n$ and $T(\mathcal{G}_n)$ the number of triangles in which a uniformly sampled node participates.

To ensure our graphs are maximally unbiased we use the maximum entropy approach from information theory [6]. This approach finds the maximally unbiased distribution as the distribution which maximizes the Shannon entropy while satisfying the given set of properties. For a simple random graph $\mathcal{G}_n$ of size n, the Shannon entropy is given by,

$$\mathcal{E}_n[\mathcal{G}_n] = -\sum_{G\in\mathbb{G}_n} \mathbb{P}\left(\mathcal{G}_n = G\right) \log(\mathbb{P}\left(\mathcal{G}_n = G\right)),$$

where $\mathbb{G}_n$ denotes the space of all simple graphs of size n.

Maximum entropy problem for graphon ensembles

Unfortunately, due to the non-linearity of the triangle constraint, solving this entropy optimization problem on graphs is hard, as was already noticed by those in the field of large deviation theory on graphs [3]. Moreover, current results are only available for dense or sub-dense graphs, while we are interested in real sparse graphs (finite average degree). We therefore translate our problem to one in the space of functional analysis, using concepts related to the theory of graph limits [2] and inhomogeneous random graphs [1]. This approach we used successfully in [4] to characterize the maximum entropy sparse graphs with given scale-free degree distributions.

To be more precise, we will consider sequences $(W_n)_{n\geq 1}$ of symmetric functions $W_n : [0,n]^2 \to [0,1]$ which we refer to as *graphon ensembles*, and let G_{W_n} be the random graph on n nodes that is created by first sampling n points $x_1, \ldots, x_n$ uniformly at random in $[0,n]$ and connecting nodes (i,j) with probability $W_n(x_i, x_j)$. We can then express $D(G_{W_n})$ and $T(G_{W_n})$ as functionals, $\mathcal{D}_n$ and $\mathcal{T}_n$ of W_n. For instance

$$\mathcal{T}_n[W_n] = \frac{\binom{n-1}{2}}{n^3} \iiint W_n(x,y)W_n(y,z)W_n(z,x)\,dx\,dy\,dz.$$

In addition, we can extent the notion of entropy to graphon ensembles by defining

$$\mathcal{E}_n[W_n] = n^{-2} \iint H(W_n(x,y))\,dx\,dy,$$

where $H(p) = -p\log(p) - (1-p)\log(1-p)$. With this we can now formulate the corresponding optimization problem for graphon ensembles.

Optimization Problem 1.1. *Let $(k_n, t_n)_{n\geq 1}$ be two sequences such that*

$$\lim_{n\to\infty} k_n = k \quad \textit{and} \quad \lim_{n\to\infty} t_n = t. \tag{2}$$

Find $(W_n^*)_{n\geq 1}$, such that for each n,

$$W_n^* = \arg\max_{W_n} \mathcal{E}[W_n],$$

while

$$\mathcal{D}_n[W_n^*] = k_n \quad \textit{and} \quad \mathcal{T}_n[W_n^*] = t_n.$$

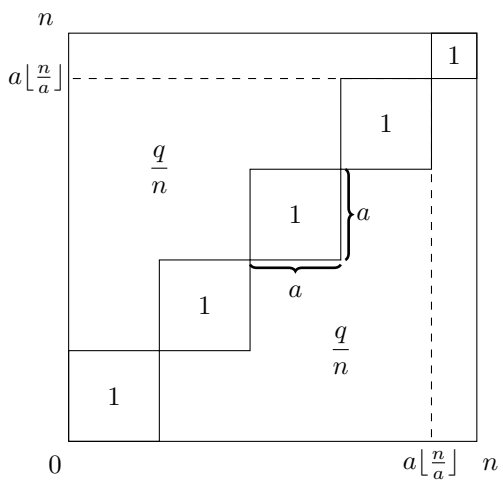


Figure 1: Representation of the symmetric Sparse Block Graphon $W_{a,q,n}$.

By solving this optimization problem we obtain a sequence of connection probability functions W_n^* that will tell us the structure of the maximum entropy random graphs with given average degree and clustering. Notice that, although we are interested in the limiting structure of W_n^*, for finite n the solution might depend on the choice of the sequences k_n and t_n. To deal with these finite effects we introduce the equivalence relation:

$$(W_n)_{n\geq 1} \equiv (W'_n)_{n\geq 1} \iff \lim_{n\to\infty} n\|W_n - W'_n\|_1 = 0,$$

where the L^1 norm is with respect to the product space of $[0, n]$ with the uniform probability measure.

Spare stochastic block graphons

To characterize the solution to Optimization Problem 1.1, let $a > 0$ and consider the partition of the interval $[0, n]$ into $\lfloor a/n \rfloor$ parts of size a and possibly one smaller part at the end. This partition induces a partition of the square $[0, n]^2$ into blocks of size $a \times a$ and possibly one smaller block. We then consider the class of Stochastic Block Graphons $W_{a,q,n}$ which are defined to be 1 on the blocks and q/n outside, see Figure 1. These graphons generated graphs which consist of $\lfloor a/n \rfloor$ cliques, of expected size a, and nodes in different cliques are connected at random with probability q/n. Our main result is the following:

Theorem 1.1. *Let $t > 0$ and $k > \sqrt{2t}$ and let k_n and t_n be any two sequences satisfying (2). Then the solution to Optimization Problem 1.1 is given, up to equivalence, by the graphon enesmble $(W_{a,q,n})_{n\geq 1}$ (see Figure 1), where $a = \sqrt{2t}$ and $q = \sqrt{2t} - k$.*

Conclusion and discussion

In this work we continued the theoretical paradigm for solving maximum entropy problems on sparse graphs using graphon ensembles [4], combining the linear constraint of average degree with the non-linear constraint of triangles. Our results show that, in the $n \to \infty$ limit, the entropy maximizing graphon ensembles with given average degree k and average triangles t is given by a Stochastic Block Model consisting of cliques whose size is Poisson Distributed with mean $\sqrt{2t}$, and which are connected at random with probability $(k - \sqrt{2t})/n$. These result shed a new light on the structures imposed by the clustering constraint. In particular we observe that the solution is not a geometric graphon as was conjectured in [5]. Therefore, it remains an open question which structural properties give rise to hidden geometry in networks.

Solving the graphon ensemble entropy problem is just one step in our research agenda. To come full circle we want to establish that the random graphs coming from this ensemble also maximize the entropy under the constraints. One approach is to establish an asymptotic relation between the graph and graphon entropy, as was done in [4]. Here, however, we encounter a new phenomenon, namely the entropy associated with the symmetry of the graphon structure. We are currently working on classifying this entropy and understanding its relation to non-linear graph structures. However, we conjecture that, due to the permutation symmetry with respect to the blocks, our Stochastic Block Graphon will remain the entropy maximizer.

References

[1] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures & Algorithms*, 31(1):3–122, 2007.

[2] C. Borgs, J. T. Chayes, H. Cohn, and N. Holden. Sparse exchangeable graphs and their limits via graphon processes. *arXiv preprint arXiv:1601.07134*, 2016.

[3] S. Chatterjee. An introduction to large deviations for random graphs. *Bulletin of the American Mathematical Society*, 53(4):617–642, 2016.

[4] P. van der Hoorn, G. Lippner, and D. Krioukov. Sparse maximum-entropy random graphs with a given power-law degree distribution. *Journal of Statistical Physics*, 2017.

[5] D. Krioukov. Clustering implies geometry in networks. *Physical review letters*, 116(20):208302, 2016.

[6] J. Shore and R. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37, 1980.

STRUCTURAL AND FUNCTIONAL REDUNDANCY IN BIOLOGICAL NETWORKS

Alice C.U. Schwarze, Mason A. Porter, Jonny Wray

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Several scholars of evolutionary biology have suggested that functional redundancy (also known as biological "degeneracy") is important for robustness of biological networks [5, 8, 9]. For networks with Ornstein–Uhlenbeck dynamics, Tononi et al. proposed measures of structural and functional redundancy that are based on mutual information between subnetworks [7]. For a network of n vertices, an exact computation of these quantities requires $O(n!)$ time. We derive expansions for these measures that one can compute in $O(n^3)$ time. We use the expansions to compare the contributions of different types of motifs to a network's structural and functional redundancy. We compute structural and functional redundancy for protein-interaction networks and find that these networks have larger functional redundancy than corresponding realisations of several random-graph models.

Structural and functional redundancy

Structural redundancy indicates the existence of structurally similar subsystems that can perform the same function. *Functional redundancy* indicates the existence of structurally different subsystems that can perform the same function [7]. Several scholars have emphasised the importance of functional redundancy to the robustness and evolvability of complex biological systems [5, 8, 9]. It is thus important to study functional redundancy in biological networks. In Fig. 1, we show an example of a metabolic network with structurally and functionally redundant subnetworks.

An information-theoretic framework

Several researchers have proposed frameworks for quantifying structural or functional redundancy in several families of networks [1, 4, 6, 7]. Tononi et al. proposed a framework for measuring structural and functional redundancy in networks with Ornstein–Uhlenbeck dynamics [2, 7]. They partitioned a network $\mathbf{X}$ into a *kernel system* $\mathbf{x}$ and an *output sheet* $\mathbf{o}$. For every k-node subsystem $\mathbf{x}_{j_k}$ of $\mathbf{x}$, they associated its function with the influence of its state on the state of $\mathbf{o}$. To quantify the influence of the state of $\mathbf{x}_{j_k}$ on the state of the output sheet, they used mutual information $\mathrm{MI}(\mathbf{x}_{j_k}, \mathbf{o})$. For a network with a kernel system of n nodes, they defined its redundancy $R(\mathbf{x}, \mathbf{o}) := \sum_{j=1}^{n} \mathrm{MI}(x_{j_1}, \mathbf{o}) - \mathrm{MI}(\mathbf{x}, \mathbf{o})$ and its functional redundancy

$$\mathrm{FR}(\mathbf{x}, \mathbf{o}) := \sum_{k=1}^{n} \left[\frac{k}{n} R(\mathbf{x}, \mathbf{o}) - \langle R(\mathbf{x}_{j_k}, \mathbf{o}) \rangle_k \right] , \qquad (1)$$

where $\langle R(\mathbf{x}_{j_k}) \rangle_k$ is the redundancy of a subsystem of $\mathbf{x}_{j_k}$, averaged over all subsystems of $\mathbf{x}$ of size k.

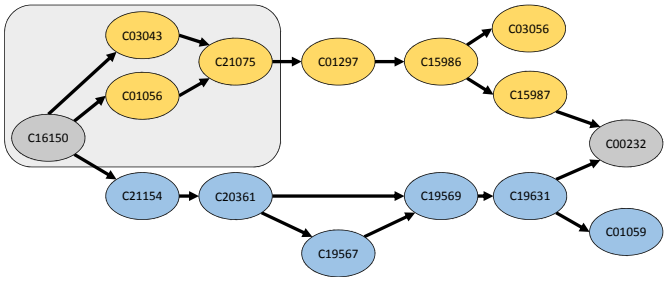


Figure 1: **Metabolic network with structurally and functionally redundant components**. Nicotine (C16150) can degrade to succinate semialdehyde (C00232) along different pathways [3]: the pyridine pathway (yellow nodes) and the pyrrolidine pathway (blue nodes). In this metabolic network for nicotine degradation, the set of yellow nodes and the set of blue nodes induce structurally different subnetworks that exhibit functional redundancy. Node labels correspond to KEGG compound identifiers [3]. In the grey box, there are two directed paths from C16150 to C21075 that are structurally redundant.

Approximation for mean functional redundancy

A problem of this approach to the calculation of functional redundancy is that the associated computation time is $O(n!)$. For many biological networks — e.g. protein–protein interaction networks and gene co-expression networks — another problem is that there is no obvious choice of output sheets. To address these problems, we focus on mean functional redundancy $\overline{\mathrm{FR}}(\mathbf{X})$, which is

the mean of $\mathrm{FR}(\mathbf{x}, \mathbf{o})$ over all possible choices of an output sheet $\mathbf{o}$ in $\mathbf{X}$. Following an approach by Barnett et al. [2] for other information-theoretic network measures, we expand $\overline{\mathrm{FR}}(\mathbf{X})$ in directed, weighted networks to a weighted sum of counts of motifs of increasing size. For undirected, unweighted networks without self-loops, this expansion simplifies to a weighted sum of counts of cycles of increasing length:

$$\overline{\mathrm{FR}}(\mathbf{X}) = \eta^2 \left(\frac{2N+1}{8}\right) \sum_{i \neq j} a_{ij}^2 + \eta^3 \left(\frac{8N-3}{4}\right) \sum_{\substack{i,j,k \\ i \neq j \neq k, i \neq k}} a_{ij} a_{jk} a_{ki} + O(\eta^4)\,, \quad (2)$$

where N is the number of nodes in $\mathbf{X}$, the matrix $\mathbf{A} = (a_{ij})$ is the adjacency matrix of $\mathbf{X}$, and $\eta \in (0, 1)$ is a normalisation factor that ensures that the sum converges.

Motifs and functional redundancy

For directed and undirected networks, we derive higher-order terms of eq. (2). We identify the motifs that contribute to $\overline{\mathrm{FR}}$. To understand their relevance to functional redundancy, we compare these motifs to motifs that contribute to covariance, correlation, and entropy in Ornstein–Uhlenbeck networks. In Fig. 2, we show the relevant motifs for covariance, correlation, entropy, and functional redundancy in directed networks. For undirected networks, our results suggest that simple cycles contribute more to functional redundancy than non-simple cycles.

One can derive an expression for mean structural redundancy $\overline{\mathrm{SR}}$ that is a weighted sum in counts of motifs of increasing size. From comparing the coefficients, we identify motifs that contribute more strongly to functional than to structural robustness.

Functional redundancy in protein-interaction networks

We compute $\overline{\mathrm{FR}}$ for protein-interaction networks and find that these networks have larger functional redundancy than corresponding realisations of several random-graph models.

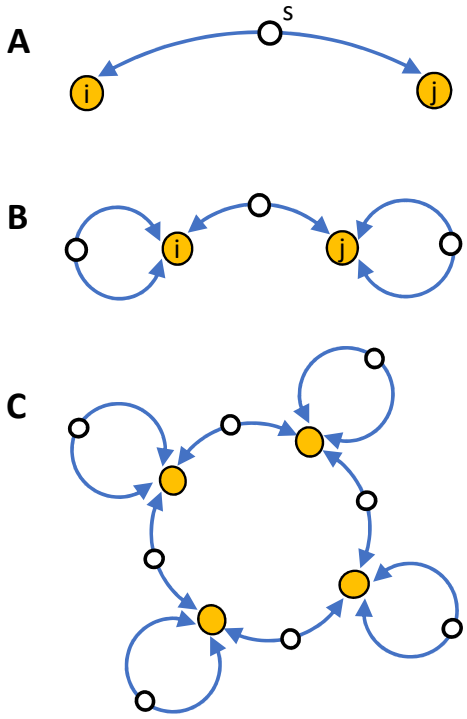


Figure 2: **Motifs that contribute to covariance, correlation, entropy, and functional redundancy in directed Ornstein–Uhlenbeck networks**. (**A**) A covariance-inducing motif (CIM) consists of two paths that connect a source node s to nodes i and j. (**B**) A correlation-inducing motif (CrIM) consists of one CIM that connects two different nodes i and j along ith several circular CIMs that connect node i to itself or node j to itself. (**C**) Motifs that contributed to entropy and functional redundancy are cycles of CrIMs.

References

[1] R. Albert, B. Dasgupta, R. Hegde, G. S. Sivanathan, A. Gitter, G. Gürsoy, P. Paul, and E. Sontag. Computationally efficient measure of topological redundancy of biological and social networks. *Physical Review E*, 84(3):1–15, 2011.

[2] L. Barnett, C. L. Buckley, and S. Bullock. Neural complexity and structural connectivity. *Physical Review E*, 79(5):051914, 2009.

[3] M. Kanehisa and S. Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, 2000.

[4] J. Macia and R. V. Solé. Distributed robustness in cellular networks: Insights from synthetic evolved circuits. *Journal of the Royal Society Interface*, 6(33):393–400, 2009.

[5] P. H. Mason, J. F. Domínguez D., B. Winter, and A. Grignolio. Hidden in plain view: Degeneracy in complex systems. *BioSystems*, 128:1–8, 2015.

[6] Z. Sun and R. Albert. Node-independent elementary signaling modes: A measure of redundancy in boolean signaling transduction networks. *Network Science*, 4(3):273–292, 2016.

[7] G. Tononi, O. Sporns, and G. M. Edelman. Measures of degeneracy and redundancy in biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 96(6):3257–3262, 1999.

[8] A. Wagner. Distributed robustness versus redundancy as causes of mutational robustness. *Bioessays*, 27(2):176–188, 2005.

[9] J. M. Whitacre. Biological robustness: Paradigms, mechanisms, and systems principles. *Frontiers in Genetics*, 3:67, 2012.

A NEW DIMENSION-REDUCTION METHOD FOR COMPLEX DYNAMICAL NETWORKS

Edward Laurence, Nicolas Doyon, Louis J. Dubé, and Patrick Desrosiers

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We introduce a new dimension-reduction method to describe the large scale behaviour of dynamical processes running on networks, primarily based on the spectral properties of the weighted adjacency matrices that characterize the interactions on the networks. The structural complexity of the networks is used to naturally set the adequate dimensionality of the reduced system. We present and compare three variants of our method. We show that our approximation scheme, even when forced to produce one-dimensional reduced systems, always gives a better description of the dynamics than the one proposed by Gao *et al.*[1].

Introduction

Dynamics of large complex networks can sometimes be modeled as simpler and lower dimensional systems. These reduced systems, if properly inferred, should provide intuitive insights about the global behaviour of the systems and help predict their dynamical resilience or breakdown.

We consider a network of N nodes encoded by the weighted and directed adjacency matrix $\boldsymbol{W} = (w_{ij})$, where the element $w_{ij} \geq 0$ indicates the strength of the directed interaction from node j to node i. Node i has an activity $x_i \in \mathbb{R}$ that evolves according to

$$\dot{x}_i = F(x_i) + \sum_{j=1}^{N} w_{ij} G(x_i, x_j). \tag{1}$$

Recent studies suggest that the global equilibrium states of such a N-dimensional system can be reduced to a one-dimensional universal function [1]. Two effective structural and activity parameters can then be extracted to describe the evolution of the system. While this is a good approximation for uncorrelated network structures [2], it fails when degree correlations become important.

Generalizing the dimension-reduction

To reduce the dimensionality of the dynamical system (1), we introduce new weighted averages that describe the global dynamics propagating on the network and the large-scale structure of the network. For each k, let $\boldsymbol{a}^{(k)} = (a_1^{(k)}, \ldots, a_N^{(k)})$ be a discrete probability distribution, i.e., $a_i^{(k)} \geq 0 \; \forall \, i$ and $\sum_i a_i^{(k)} = 1$. These probability distributions allow to define weighted average activities, $\langle x \rangle^{(k)} = \sum_{j=1}^{N} a_j^{(k)} x_j$, as well as the weighted average in-degrees, $\langle w \rangle^{(k)} = \sum_{j=1}^{N} a_j^{(k)} w_j^{\text{in}}$.

We can show that the dynamics of the weighted average activities $\langle x \rangle^{(k)}$, describing the system (1), is approximately governed by the following system:

$$\begin{aligned}
\dot{\langle x \rangle}^{(0)} &= F(\langle x \rangle^{(0)}) + \langle w \rangle^{(0)} G(\langle x \rangle^{(1)}, \langle x \rangle^{(1)}) \\
\dot{\langle x \rangle}^{(1)} &= F(\langle x \rangle^{(1)}) + \langle w \rangle^{(1)} G(\langle x \rangle^{(2)}, \langle x \rangle^{(2)}) \\
&\vdots
\end{aligned}$$

where the distributions are transformed according to

$$\langle w \rangle^{(k)} \boldsymbol{a}^{(k+1)} = \boldsymbol{W}^T \boldsymbol{a}^{(k)}. \tag{2}$$

Transformation (2) is a well-defined map on the space of probability distributions since it preserves both normalization (i.e., $\sum_i a_i^{(k+1)} = 1$ if $\sum_i a_i^{(k)} = 1$) and positiveness (i.e., $a_i^{(k+1)} \geq 0$ if $a_i^{(k)} \geq 0$).

At first glance, the new dynamical system is unsolvable; it contains an infinite number of dynamical variables. However, an appropriate choice of the initial probability distribution $\boldsymbol{a}^{(0)}$ enables us to close the set of differential equations and get a d-dimensional system. We have developed three procedures to choose $\boldsymbol{a}^{(0)}$, all of them leading to a d–dimensional reduced system: (i) the d–period method (ii) the power iteration method and (iii) the eigenvector composition method.

The d–period method consists in choosing $\boldsymbol{a}^{(0)}$ such that $\boldsymbol{a}^{(d+1)} = \boldsymbol{a}^{(0)}$. In doing so, the d-th differential equation closes the system since $\dot{\langle x \rangle}^{(d)} = f(\langle x \rangle^{(d)}, \langle x \rangle^{(0)})$. We achieve this by letting $\boldsymbol{a}^{(0)}$ be the positive eigenvector of the d-th power of $\boldsymbol{W}^T$.

We can show that the d–period method always works for strongly connected networks, i.e., networks in which there is a path between each pair of nodes. The case $d = 1$ is of particular interest since it always leads to a one-dimensional reduced system in which the distribution $\boldsymbol{a}^{(0)}$

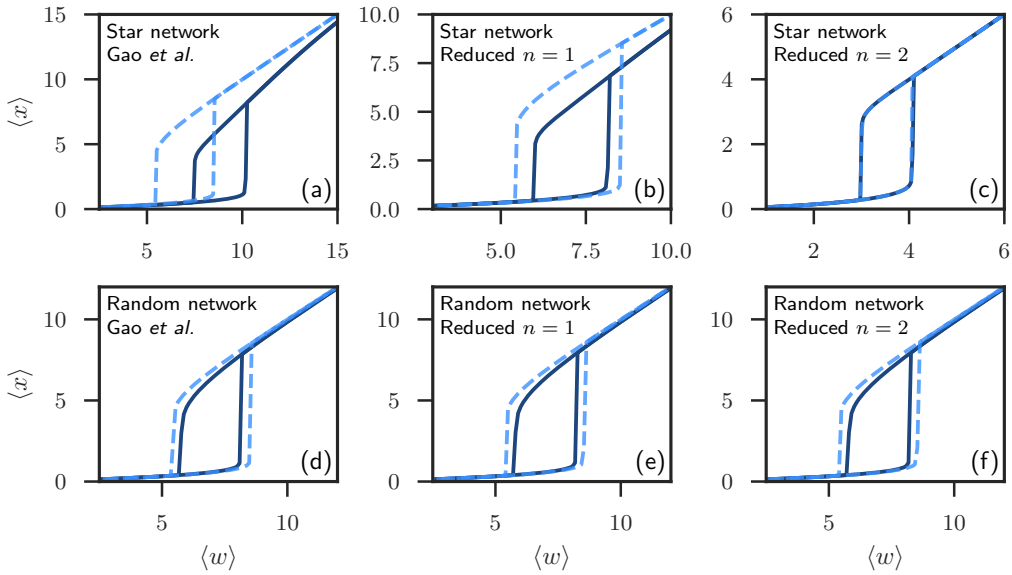


Figure 1: Weighted average activity at equilibrium $\langle x \rangle$ as a function of the weighted average input degree $\langle w \rangle$ using Gao et al. reduction (first column) and our approach with $n = 1$ and $n = 2$ global variables (second and third columns) for star networks of 6 nodes and Gilbert random graph of 100 nodes of density $p = 0.1$. Dashed lines are the predictions of each approach while full lines are obtained from numerical simulations on networks with neural dynamics.

is identical to the eigenvector centrality. The formalism proposed by Gao *et al.* is an approximation of the $(d = 1)$–period method.

In the power iteration method, we set $\boldsymbol{a}^{(0)}$ as the uniform probability distribution, i.e., $a_i^{(0)} = 1/N$ for all i. Then, as we apply the transformation (2), $\boldsymbol{a}^{(k)}$ aligns with the dominant eigenvector of $\boldsymbol{W}^T$. When the convergence $\|\boldsymbol{a}^{(k+1)} - \boldsymbol{a}^{(k)}\| < \epsilon$ is reached, the k-th equation of the reduced system can safely be approximated by $\langle \dot{x} \rangle^{(k)} = f(\langle x \rangle^{(k)})$. In practice, this method should be used only to obtain the evolution of the uniform activity average.

The eigenvector composition method consists in choosing $\boldsymbol{a}^{(0)}$ as a linear composition of d dominant and linearly independent eigenvectors of $\boldsymbol{W}^T$. In doing so, $\boldsymbol{a}^{(d)} = \sum_{j=0}^{d-1} c_j \boldsymbol{a}^{(j)}$ is simply a linear composition of the constructed probability distribution.

Results

We apply our formalism to different network structures (Fig. 1) on which the activity evolves according to a well-known dynamics in computational neuroscience [3],

$$\dot{x}_i = -x_i + \sum_{j=1}^{N} w_{ij} \sigma(x_j - \mu)$$

where μ is a parameter and $\sigma(\cdot)$ is the sigmoid function. For random networks, the 1–period method performs as well as the Gao *et al.* formalism. As expected by the spectral analysis of the weighted adjacency matrix, higher dimensions of the reduction do not improve the description.

An impressive demonstration of the power of our formalism is given by star graphs where a single core node is connected to $N-1$ periphery nodes. The one-dimensional system of Gao *et al.* overestimates the activation of the core node so that large discrepancies are visible, even for small graphs $N = 6$, and increase with the size of the graph. Using our formalism, we describe *exactly* the star graph using a two-dimensional system. Moreover, we show that $\langle x \rangle^{(0)}$ and $\langle x \rangle^{(1)}$ describe the activity of the core node and the periphery nodes respectively.

References

[1] J. Gao, B. Barzel, and A.-L. Barabási. Universal resilience patterns in complex networks. *Nature*, 530(7590):307, 2016.

[2] C. Tu, J. Grilli, F. Schuessler, and S. Suweis. Collapse of resilience patterns in generalized Lotka-Volterra dynamics and beyond. *Phys. Rev. E*, 95(6):062307, 2017.

[3] H. R. Wilson and J. D. Cowan. Excitatory and inhibitory interactions in localized populations of model neurons. *Biophys. J.*, 12(1):1–24, 1972.

ESTIMATING EPIDEMIC ARRIVAL TIMES USING LINEAR SPREADING THEORY

Lawrence M. Chen, Matt Holzer, Anne Shapiro

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Modern transportation networks allow for the rapid global spread of disease. In this context, a fundamental property of interest are arrival times. Given a disease that originates in a particular city or location arrival times describe how long it takes for that disease to arise in some other city, We develop methods to estimate arrival times based upon the linear determinacy of the system and the theory of traveling fronts.

Traditional mathematical analysis of the spread of disease has focused on partial differential equation (PDE) models and the study of traveling waves. Ostensibly, the prevalence of long range connections between cities via airline travel would render this approach ineffective. On the contrary, we demonstrate that when the rate of diffusion is small relative to the reaction rates then the dynamics are still dominated by the formation of traveling fronts only with a different notion of distance between cities to accommodate the long range connections; see [1, 3, 5].

Model Description

We study the following meta-population susceptible-infected-reduced (SIR) model described in [1]:

$$\begin{aligned}
\partial_t s_n &= -\alpha s_n j_n + \gamma \sum_{m\neq n} P_{nm}(s_m - s_n) \\
\partial_t j_n &= \alpha s_n j_n - \beta j_n + \gamma \sum_{m\neq n} P_{nm}(j_m - j_n) \\
\partial_t r_n &= \beta j_n + \gamma \sum_{m\neq n} P_{nm}(r_m - r_n). \qquad (1)
\end{aligned}$$

Here s_n, j_n and r_n are the normalized susceptible, infected and recovered populations of city n, respectively. The parameter α describes the rate of infection and β is the mean recovery rate of individuals. The parameter γ is the average mobility rate, meaning the percentage of the population of the system traveling at any given time. The matrix P is a weighted adjacency matrix with $0 \leq P_{nm} \leq 1$. It is row stochastic, with each P_{nm} corresponding to the probability of a random walker at node n moving to node m.

System (1) is a reaction-diffusion equation on a network. For this study, we consider the worldwide airline transportation network studied in [4]. We will focus on arrival times, meaning the first time at which a disease is established in a certain city. Mathematically, for a disease initialized in city n we define the arrival time at city m as

$$T_{nm} = \inf\{\ t \geq 0 \mid j_m(t) = \kappa\ \},$$

for some threshold $0 < \kappa \ll 1$.

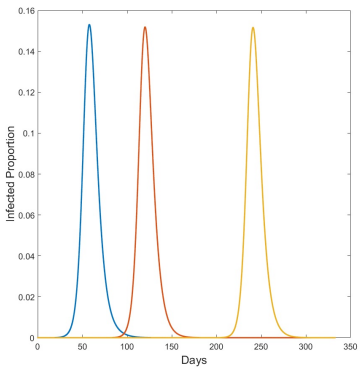


Figure 1: Proportion of infected population in Paris, France (blue), Kodiak, Alaska (Red), and Wasu, Papua New Guinea (Yellow) as a function of time with infection originating in Paris. Here $\alpha = 0.5$, $\beta = 0.25$ and $\gamma = 0.001$

Main Outcomes

In the PDE context, fronts speeds for the SIR model are *linearly determined* meaning that they can be calculated using only information about the linearization of the PDE at the unstable disease free state. The network dynamical system (1) enjoys a similar property and we utilize this to predict arrival times in the nonlinear system using the linearized system.

Numerical simulations of (1) for small values of γ reveal a general linear trend between observed arrival times and distance – in terms of number of flights – between cities; see Figure 2. Nonetheless, even for very small γ values significant differences in arrival times are observed amongst cities a fixed number of flights for the origination city.

Our main contribution is the derivation of an explicit estimate for disease arrival times based upon expansion of the heat kernel of the Graph Laplacian; see [2]. A comparison of predicted arrival times versus numerically observed times are depicted in Figure 2. We outline that derivation below.

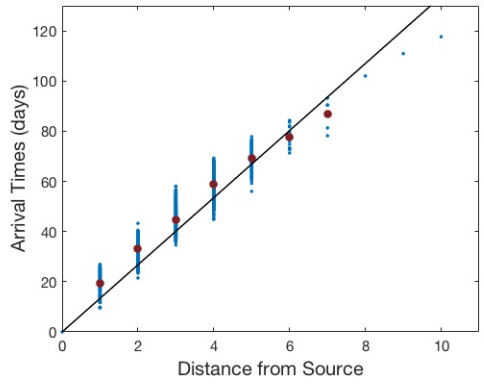


Figure 2: Arrival times compared to distance from infection source nodewith $\gamma = 0.01$ for an infection originating in Paris, France.

To estimate the linear arrival times, we linearize (1) near the unstable disease free equilibrium and find that the evolution for the infected population decouples. For a disease originating in city n with initial infected proportion j_0, the infected proportion at city m under the linearized dynamics is,

$$j_m(t) = j_0 v_m^T e^{\Gamma t} e^{\gamma \mathrm{P} t} v_n,$$

Expand the matrix exponential $e^{\gamma \mathrm{P} t}$ as a series

$$v_m^T e^{\gamma \mathrm{P} t} v_n = \sum_{k=0}^{\infty} \frac{\gamma^k t^k}{k!} v_m^T \mathrm{P}^k v_n. \tag{2}$$

Let $\rho_k = v_m^T \mathrm{P}^k v_n$. Note that ρ_k describes the probability of a random walker starting at city m being located in city n after k steps. Suppose that the minimal number of flights required to travel from city n to city m is d. Then, all terms before the dth in the expansion are zero. Assuming that the leading order term dominates, we solve

$$j_0 e^{\Gamma t} \frac{\gamma^d \rho_d t^d}{d!} = \kappa,$$

and obtain the heat kernel estimate for the arrival time,

$$T_{nm}^{HK} = \frac{d}{\alpha - \beta - \gamma} W\left(\frac{(d!)^{1/d}}{d} \frac{\alpha - \beta - \gamma}{\gamma (\rho_d)^{1/d}} \left(\frac{\kappa}{j_0} \right)^{1/d} \right), \tag{3}$$

where W is the Lambert-W function.

A comparison of theoretical arrival times versus those observed in numerical simulations of (1) is depicted in Figure 3.

The estimate (3) is accurate for asymptotically small values of the diffusion parameter γ. For moderate to large values of γ, accurate predictions can again be recovered by including more terms from the heat kernel expansion.

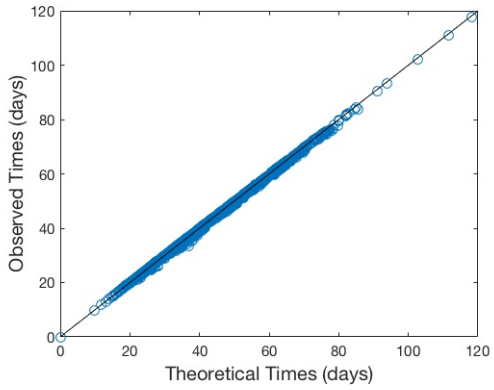


Figure 3: Observed arrival times compared to predictions using the heat kernel expansion (3) for the data depicted in Figure 2.

References

[1] D. Brockmann and D. Helbing. The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342, 2013.

[2] L. M. Chen, M. Holzer, and A. Shapiro. Estimating epidemic arrival times using linear spreading theory. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 28(1):013105, 2018.

[3] A. Gautreau, A. Barrat, and M. Barthélemy. Global disease spread: Statistics and estimation of arrival times. *Journal of Theoretical Biology*, 251(3):509 – 522, 2008.

[4] R. Guimerà, S. Mossa, A. Turtschi, and L. A. N. Amaral. From the Cover: The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proceedings of the National Academy of Science*, 102:7794–7799, May 2005.

[5] F. Iannelli, A. Koher, D. Brockmann, P. Hövel, and I. M. Sokolov. Effective distances for epidemics spreading on complex networks. *Phys. Rev. E*, 95:012313, Jan 2017.

A BRIDGE BETWEEN HOMOTOPY THEORY AND NETWORK SCIENCE

Leo Torres, Pablo Suárez Serrato and Tina Eliassi-Rad

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Abstract

Nonbacktracking cycles (NBCs) have recently gained popularity in the Network Science community. We highlight the connections with homotopy theory from algebraic topology and show how NBCs capture structural measures such as degree distribution and clustering coefficient. We propose a graph distance measure based on NBCs that is well grounded in results from homotopy theory.

NBCs and Homotopy Theory

Homotopy theory is a branch of algebraic topology, alongside (persistent) homology. Homotopy studies algebraic invariants of topological spaces called *homotopy groups*. Given a graph or simplicial complex, its first homotopy group is called *the fundamental group* and denoted by π_1. The relationship between homotopy and homology groups is well known:[1] for example, the first homology group is the abelianization (i.e., a reduced version) of π_1. Regardless, homology groups are often preferred in practical applications because they are easier to compute in general. However, in the case of graphs, studying π_1 is no more difficult than analyzing the so-called *nonbacktracking matrix* B, as we discuss below.

The nonbacktracking matrix B has been the focus of recent works in relation to centrality [1], modularity [2], and embedding [3]. Given a graph with m edges, B is the $2m \times 2m$ matrix where each edge is represented by two rows and two columns, one per orientation: (u, v) and (v, u). For two edges (u, v) and (k, l), B is given by

$$B_{u\to v,k\to l} = \delta_{vk}(1 - \delta_{ul}),$$

where δ_{ij} is the Kronecker delta. Thus, there is a 1 in the entry indexed by row (u, v) and column (k, l) when $u \neq l$ and $v = k$; and a 0 otherwise. To our knowledge, B has been studied *independently* of its connections to homotopy, and thus its full potential as a bridge between topology and network science has yet to be exploited.

Concretely, the connection to homotopy is realized as follows. Let $G = (V, E)$ be a graph. For $e = (u, v) \in E$, define e^{-1} as the same edge traversed in the inverse order, $e^{-1} = (v, u)$. A *cycle* in G is a sequence of edges $e_1e_2...e_k$ such that if $e_i = (u_i, v_i)$ then $v_i = u_{i+1}$ for $i = 1, ..., k-1$ and $v_k = u_1$. Here, k is called *length* of the cycle. A *nonbacktracking cycle* (NBC) is one where $e_{i+1} \neq e_i^{-1}$, $i = 1, ..., k-1$ and $e_k \neq e_1^{-1}$; that is, an edge is never followed by its own inverse. Terras [4] showed that the set of NBCs is in one-to-one correspondence with the set of conjugacy classes of π_1.[2] In a nutshell, this is because backtracking edges are homotopically trivial. Now, the matrix B tracks each pair of incident edges that do not comprise a backtrack. Thus, we can study the set of NBCs, and B, as a proxy for studying π_1.

Here, we propose a way to take advantage of the connection between homotopy and the nonbacktracking matrix B. It is known that the so-called *length spectrum*, a function defined on π_1, is particularly apt at distinguishing graphs, even up to isometry [5, 6]; and describes many graph properties such as girth, number of spanning trees of G, and whether it is a forest, bipartite, or regular, among others [7]. The length spectrum assigns to each cycle its length k. However, for a graph G, the trace of the kth power of B gives the number of NBCs of length k. Thus, in the this work we focus on the question: **is the length spectrum, seen through the eigenvalues of the nonbacktracking matrix, appropriate for network science and its applications?** We answer this question in the positive by showing that B tracks important network measures such as the degree distribution and clustering coefficient. We also show an efficient algorithm to compute B as well as explain geometric patterns found in the distribution of B's eigenvalues. Finally, We propose a method for computing graph distance based on the length spectrum, i.e., using the eigenvalues of B.

[1] This result is known as the Hurewicz theorem. https://en.wikipedia.org/wiki/Hurewicz_theorem

[2] Conjugation is an operation over the elements of a group that defines an equivalence relation. The equivalence classes are called conjugacy classes. https://en.wikipedia.org/wiki/Conjugacy_class

NBCs in B and its Uses in Network Science

We focus on information contained in NBCs and B. Note that we do not propose B to be superior to other matrix representations in all cases. Rather, we show that NBCs in B carry information that is relevant and useful in studying complex networks. To our knowledge, this information has not been utilized before.

Here is our algorithm for computing B. Define the $n \times 2m$ incidence matrices $M^+_{x,u\to v} = \delta_{xu}$ and $M^-_{x,u\to v} = \delta_{xv}$, and write $C = (M^+)^T M^-$. Therefore,

$$B_{u\to v,k\to l} = C_{u\to v,k\to l}(1 - C_{u\to v,k\to l})$$

Thus, we can compute B in a single pass over the nonzero entries of C. For networks with no degree correlations, we find $nnz(C) = O(n\langle k^2\rangle)$, where n is the number of nodes, $\langle k^2\rangle$ is the second moment of the degree distribution, and $nnz(C)$ measures the number of non-zero entires in the matrix C. Since computing M^+, M^- takes $O(m)$ time, we can compute B in time complexity $O(m+n\langle k^2\rangle)$. Observe that the sparsity of B also grows with the second moment of the degree distribution, $nnz(B) = O(n\langle k^2\rangle)$. Contrast this to the fact that the sparsity of the adjacency matrix A grows with the first moment of the degree distribution, $nnz(A) = O(n\langle k\rangle)$. In the case of a power-law degree distribution with exponent $2 \le \gamma \le 3$, the runtime of our algorithm falls between $O(m+n)$ and $O(m+n^2)$, and we have $\log nnz(B) \le O(\frac{1}{\gamma-1})$.

Next, we turn to B's eigenvalues and their relation to the number of triangles. Write $\lambda_k = a_k + ib_k \in \mathbb{C}$ for the complex eigenvalues of B. The number of triangles in a network is proportional to $tr(B^3) = \sum_k a_k(a_k^2 - 3b_k^2)$. On the one hand, B's eigenvalues tend to fall on a circle in the complex plane. On the other hand, if $\sum_k a_k^2$ is large and $\sum_k b_k^2$ is small (implying a large number of triangles), the λ_k cannot all fall in the circle: the more triangles, the less marked the circular shape. (Figure 1a.)

Application: Graph Distance

Given two graphs G_1, G_2 and a positive integer r, compute the top r eigenvalues of their corresponding nonbacktracking matrices. Write these as $\lambda_k = a_k + ib_k$ for G_1 and $\mu_k = \alpha_k + i\beta_k$ for G_2. Assign to G_1 the feature vector $v_1 = (\mathbf{a}, \mathbf{b}) = (a_1, ..., a_r, b_1, ..., b_r)$ and $v_2 = (\boldsymbol{\alpha}, \boldsymbol{\beta})$ to G_2. Define the distance score of G_1 and G_2 as $\|v_1 - v_2\|$. (See Figure 1b.) Preliminary results show that both low and high values of r yield similar performance in a clustering setting, while $r \sim 15$ becomes optimal. Thus, choosing an appropriate value of r is a non-trivial task that will require further research. One advantage of this distance is that it can be fine tuned to capture certain features; e.g., if clustering is of particular interest, one may use the modified feature vectors $(\sigma\mathbf{a}, \mathbf{b}/\sigma)$ and $(\sigma\boldsymbol{\alpha}, \boldsymbol{\beta}/\sigma)$, for some $\sigma \ge 1$.

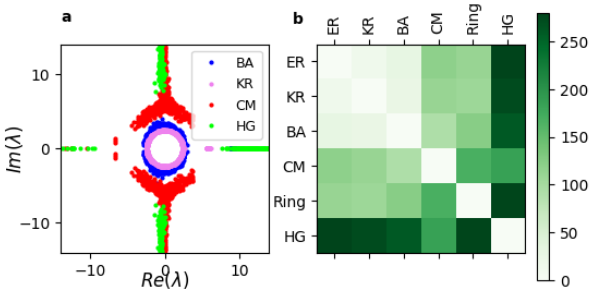


Figure 1: **a)** Eigenvalues of B of random graph models. Graphs with more triangles form a less marked circular shape around the origin. **b)** Average distance between several random graph models. Graphs are sorted in ascending number of triangles from left to right. ER: Erdos-Renyi, BA: Barabasi-Albert, KR: Kronecker Graph, CM: Configuration Model, HG: Hyperbolic Graph, Ring: ring lattice.

Acknowledgements

We thank Evimaria Terzi for her contributions. Torres and Eliassi-Rad were supported by NSF CNS-1314603, NSF IIS-1741197, and DTRA HDTRA1-10-1-0120.

References

[1] Martin, Travis, et al. *Localization and Centrality in Networks.* Phy. Rev. E. 90(5), Dec. 2014.

[2] Krzakala, F., et al. *Spectral Redemption in Clustering Sparse Networks.* PNAS 110(52):20935-20940, 2013.

[3] Jiang, Fei, et al. *On Spectral Graph Embedding: A Non-Backtracking Perspective and Graph Approximation.* In SIAM SDM, 2018.

[4] Terras, Audrey. "Ihara Zeta Function", in *Zeta Functions of Graphs: a Stroll through the Garden.* Cambridge University Press, 2011, pp. 11-16.

[5] Durfee, Christina, and Kimball Martin. *Distinguishing Graphs with Zeta Functions and Generalized Spectra.* Lin. Alg. and Its Appl. 481:54-82, 2015.

[6] Constantine, David, and Jean-Franois Lafont. *Marked Length Rigidity for One-Dimensional Spaces.* Journal of Topology and Analysis, 2018, pp. 137.

[7] Yaim Cooper, *Properties determined by the Ihara zeta function of a graph,* Electron. J. Combin. 16(1), 2009.

Invited Talk II

Understanding graphs through spectral densities

David Bindel

Cornell University

In this talk, we report ongoing work on the analysis of graphs via global summaries of the eigenvalue distributions and eigenvector behavior. Our approach is drawn from the condensed matter physics literature, where the idea of local and global densities of states is often used to understand the electronic structure of systems, and we describe how these densities play a common role in such seemingly disparate topics as spectral geometry, condensed matter physics, and the study of centrality measures in graphs. We then discuss how structural motifs manifest in the spectrum, give fast algorithms to estimate spectral densities, and conclude with a discussion of some of our current research directions in applying these tools to the analysis of large-scale graphs.

GRAPH REDUCTION BY EDGE DELETION AND EDGE CONTRACTION

Gecia Bravo Hermsdorff[*] *& Lee Gunderson*[*]

[*] *joint first authors*

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary: graph reduction beyond edge deletion

How might one "coarse-grain" a graph? That is, generate a reduced graph that preserves the global structure at the expense of discarding local details? There is a large body of literature on the techniques of graph sparsification, i.e. deleting a subset of the edges and reweighting the rest so as to approximately preserve, e.g. the Laplacian quadratic form [3]. Interestingly, for a planar graph, edge deletion corresponds to edge contraction in its planar dual (and more generally, for a graphical matroid and its dual). This duality suggests a way to further reduce a graph. Indeed, with respect to the dynamics induced by the Laplacian (e.g. diffusion/random walks, electrical/potential flows), deletion and contraction are physical manifestations of two opposite limits: edge weights of 0 and ∞, respectively. In this work, we propose a measure of edge importance that captures both of these operations. Based on this measure, we provide a unifying framework by which one can systematically reduce a graph while preserving its large-scale structure, not only in the number of edges, but also in the number of nodes.

Motivation: reduced computational cost and structural insights

Suppose you are playing a complicated social game in real time (e.g. living), in which anyone's decision depends (in varying degrees) on the decisions of other people. As it is difficult to store and reason about this entire social network, an efficient strategy requires constructing a reduced approximation of it with only a few essential actors and interactions. In addition, an appropriate "coarse-graining" of a network can provide structural insights.[1]

[1]For example, a common theme of many biological systems is the presence of complicated pathways that produce a relatively simple result (e.g., protein activation pathways). Semantic understanding comes from the reduction of these subsystems to their resulting behavior (e.g., X activates a chain that eventually inhibits Y).

Background: spectral sparsification by edge deletion

Given a graph $G = (V, E, w)$ (where V is the set of vertices, $E \subseteq V \times V$ is the set of edges, and $w_e \in (0, \infty)\ \forall e \in E$ are the edge weights), what does it mean for a reduced graph $\widetilde{G}$ to preserve the "structure" of G? A prototypical example of such a reduction is approximating a complete graph by a d-regular expander on the same vertex set, which effectively mimics the connectedness of the complete graph. Indeed, the spectrum of the graph Laplacian L is frequently used to characterize graph connectivity (e.g. Cheeger inequality).

This inspired Spielman and collaborators to introduce the notion of spectral approximation, where a graph $\widetilde{G}$ is an ϵ-approximation of G, if $\widetilde{G}$ preserves the Laplacian quadratic form $x^T L_G x$ within a factor of $1 \pm \epsilon$. They showed that *every* graph G has a sparse spectral approximation that can be found in nearly linear time [2, 4]. Their algorithm involves computing effective resistances to obtain a measure of edge importance, suppressing the deletion of spectrally important edges, and encouraging instead the downsampling of more redundant edges.

Contribution: a new notion of edge importance

Sparsification by sampling in this manner can be seen as a probabilistic reweighting of the edges, with some fraction of edges changed to zero weight, while others have their weight increased, such that $\mathbb{E}[L_{\widetilde{G}}] = L_G$. This reduces the number of parameters (edge weights w_e) by fixing a subset of them to a boundary, namely the edge deletion boundary $w \to 0$. In this work, we consider the opposite boundary: edge contraction $w \to \infty$.[2]

Many simulations of processes on graphs involve solving $Lx = b$ for x [5]. This suggests that one should attempt to preserve L^{-1} (or more appropriately, the pseudoinverse $L^\dagger$). Consider how the inverse Laplacian changes as we vary the weight w_e of a single edge; the Sherman-Morrison

[2]This interpretation can be viewed as a manifestation of the ideas behind the Manifold Boundary Approximation Method [6].

formula gives:

$$L^\dagger(w_e + w'_e) = L^\dagger(w_e) - \frac{w'_e}{1 + w'_e b_e^\top L^\dagger b_e} L^\dagger b_e b_e^\top L^\dagger \quad (1)$$

where b_e is the signed incidence vector for edge e. Thus, a probabilistic reweight of this edge can be chosen such that $\mathbb{E}[L^\dagger(w_e + w'_e)] = L^\dagger(w_e)$. With this goal, we define our measure of edge importance as:

$$h_e = \left\| \frac{dL^\dagger}{d \ln w_e} \right\|_{2,2}. \quad (2)$$

Our reduction algorithm samples the edges to be reduced with probability $p_e \propto 1/h_e$, appropriately reweighting the remaining ones so as to preserve $\mathbb{E}[L^\dagger_{\widetilde{G}}]$ to lowest order,[3] while attempting to minimize the variance of $L^\dagger_{\widetilde{G}} - L^\dagger_G$.

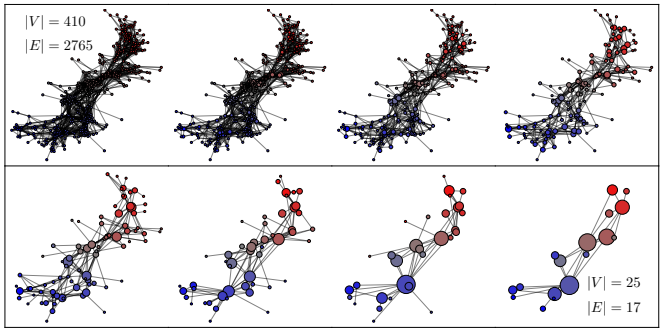


Figure 1: Our graph reduction method applied to social interaction data [1]. Color indicates lowest nontrivial eigenvector of the Laplacian.

A dual reduction: edge contraction

The limit of $w \to \infty$ is a peculiar limit with respect to the Laplacian; while the entries in L associated with the edge diverge, $L^\dagger$ remains finite; its limit being an $n \times n$ matrix whose rank is reduced by one. In fact, the rows and columns for the contracted nodes become identical, and we can write $L^\dagger$ as an $(n-1) \times (n-1)$ matrix, where the two nodes are represented by a single index, representing their coalescence into a single object.

Combining nodes means that we have to assign a weight to them, according to the "amount" of nodes they represent. Let C be the $|V'| \times |V|$ projection matrix, mapping from the original nodes V to the contracted nodes V'. Let B be the $|E| \times |V|$ signed incidence matrix, and W_e (W_n) be the diagonal matrix of edge (node) weights. The familiar graph Laplacian is given by $L = B^\top W_e B$, but this

[3] If multiple edges are reduced at once, one must use the (more general) Woodbury matrix identity.

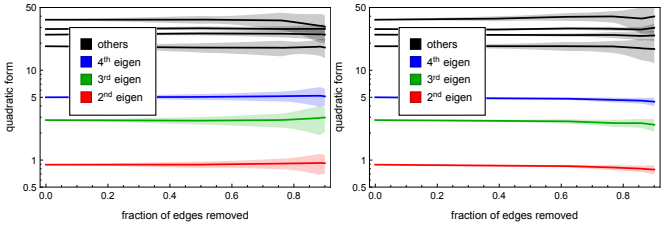


Figure 2: Quadratic form $x^\top L_{\widetilde{G}} x$ for eigenfunctions of the original Laplacian. Same initial graph as in figure 3. **Left:** Spielman and collaborators' sparsification method. **Right:** Our method, using only edge deletion.

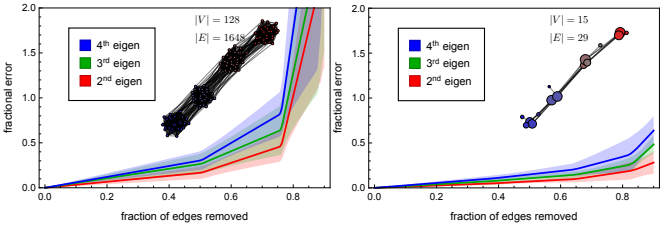


Figure 3: Fractional error in the inverse Laplacian, $\|L^\dagger_{\widetilde{G}} x - L^\dagger_G x\|_{2,2} / \|L^\dagger_G x\|_{2,2}$. **Left:** Spielman and collaborators' sparsification method. **Right:** Our method, using both deletion and contraction. **Insets:** Initial graph (stochastic line model, n=128, k=4, a=0.75, b=0.05) and reduced graph.

tacitly assumes that the nodes are identically important. Differential geometry offers a prescription for how to incorporate node weights into the Laplacian; treating a graph as a simplicial complex, the Hodge Laplacian for 0-forms (functions on vertices) is given by δd, where the differential $d = B$, and the codifferential $\delta = W_n^{-1} B^\top W_e$. Thus, in cases where the nodes have an additive measure of importance, it is appropriate to use $L = W_n^{-1} B^\top W_e B$.

References

[1] I. Lorenzo, et al. What's in a Crowd? Analysis of Face-to-Face Behavioral Networks. *J. of Theoretical Biology*, 271(1):166–180, 2011

[2] D. A. Spielman & N. Srivastava. Graph Sparsification by Effective Resistance. *SIAM Journal on Computing*, 40(6):1913–1926, 2011

[3] D. A. Spielman & S. Teng. Nearly-Linear Time Algorithms for Graph Partitioning, Graph Sparsification, and Solving Linear Systems. *arXiv:cs/0310051*, 2003

[4] D. A. Spielman & S. Teng. Spectral Sparsification of Graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011

[5] S. Teng. The Laplacian paradigm: Emerging algorithms for massive graphs. International Conference on Theory and Applications of Models of Computation 2010, 2–14

[6] M. K. Transtrum, B. B. Machta, & J. P. Sethna. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Physical Review E*, 83(3):036701, 2011

A UNIFYING FRAMEWORK FOR GRAPH CLUSTERING

Nate Veldt, David F. Gleich, Anthony Wirth

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We present LambdaCC: a new framework for community detection in networks that unifies and generalizes a number of other well-studied objectives including modularity clustering, cluster deletion, and sparsest cut. The key to LambdaCC is a resolution parameter λ, which we prove implicitly controls both the internal density and external connectivity of clusters formed by optimizing our objective. Varying λ provides a way for practitioners to design a clustering objective function that is specifically tailored to an application where there is an a priori understanding of the desired structure of output clusters.

Introduction and Background

Given a graph G with nodes V and edge set E, graph clustering or *community detection* is the task of partitioning V into disjoint groups of nodes that are more densely connected to each other than the rest of the graph. Many formal objective functions have been introduced and studied, depending on context and application. In our work we introduce a new objective LambdaCC, which generalizes and interpolates between several other objective functions that previously were not known to be related.

One popular clustering objective among computer scientists is the sparsest cut problem, which seeks the set $S \subset V$ which minimizes

$$\phi(S) = \text{cut}(S)/|S| + \text{cut}(S)/|V \backslash S|$$

where $\text{cut}(S)$ counts the number of edges between S and $V \backslash S$. Other standards of clustering quality put a greater emphasis on the internal density of clusters, such as the *cluster deletion* objective, which partitions a graph into disjoint cliques by removing the fewest edges possible. Arguably the most widely used multi-cluster objective for community detection is modularity, introduced by Newman and Girvan [2], which seeks to maximize the following quality function over arbitrary clusterings $\mathcal{C}$:

$$\mathcal{M}(\mathcal{C}) = \frac{1}{2|E|} \sum_{i \neq j} (A_{ij} - P_{ij})\, \delta_{ij}. \tag{1}$$

In the above, $A_{ij} = 1$ if nodes i and j share an edge and is zero otherwise, δ_{ij} is a zero-one indicator for whether i, j are in the same cluster, and P_{ij} represents the probability that i, j share an edge in a specified random graph null model. A closely related objective function is the Hamiltonian [3], which includes a resolution parameter γ:

$$\mathcal{H}(\mathcal{C}) = -\sum_{i \neq j} (A_{ij} - \gamma P_{ij})\, \delta_{ij}. \tag{2}$$

When $\gamma = 1$, maximizing (1) is equivalent to minimizing (2). Varying γ changes the penalty for clustering non-adjecent nodes together or separating nodes which share an edge.

One downside to modularity and the Hamiltonian is that there are no approximation guarantees for either: all current algorithms are heuristics. The clustering framework we introduce is closely related to these objectives but is more amenable to approximation results, and additionally leads to new connections between sparsest cut and the cluster deletion problem.

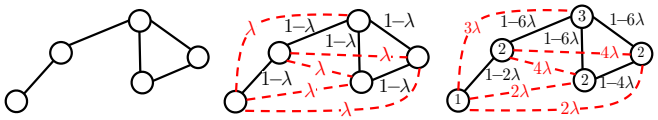


Figure 1: We convert a toy graph (left) into a signed graph for standard (middle) and degree-weighted (right) LambdaCC. Dashed red lines indicate negative edges.

Main Results

The LambdaCC framework takes an unsigned graph $G = (V, E)$ and converts it into a signed graph $G' = (V, E^+, E^-)$ using a fixed clustering resolution parameter $\lambda \in (0, 1)$. We then partition V by applying correlation clustering [1] on G', which seeks a clustering that minimizes the total weight of negative edges inside clusters plus the weight of positive edges crossing between clusters. To construct G', we first define a node weight w_v for each $v \in V$. We draw a positive edge between nodes u, v with weight $(1 - \lambda w_u w_v)$ if $(u, v) \in E$, and draw a negative edge of weight $\lambda w_u w_v$ if $(u, v) \notin E$. When $w_v = 1$ for all $v \in V$ we call this *standard* LambdaCC, and if we set w_v to be the degree d_v of $v \in V$, we call this *degree-weighted*. Figure 1 illustrates the process of constructing G'.

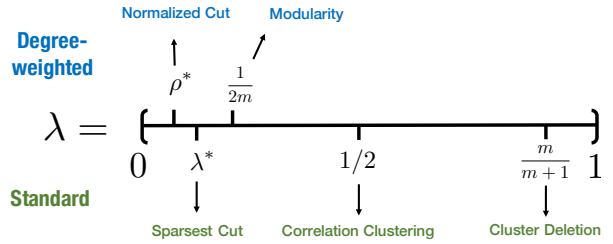


Figure 2: LambdaCC is equivalent to several other objectives for specific values of $\lambda \in (0, 1)$. Normalized cut is a degree-weighted analog to sparsest cut.

The LambdaCC objective can be formally expressed in terms of edges and non-edges in G:

$$\lambda cc(\mathcal{C}) = \sum_{(i,j)\in E} (1-\lambda w_i w_j)(1-\delta_{ij}) + \sum_{(i,j)\notin E} \lambda w_i w_j \delta_{ij} \quad (3)$$

By setting $P_{ij} = d_i d_j/(2|E|)$ and $\gamma = 2|E|\lambda$, we the following relationship:

$$\lambda cc(\mathcal{C}) = \textstyle\sum_{(i,j)\in E}(1 - \lambda d_i d_j) + \mathcal{H}(\mathcal{C})/2\,, \quad (4)$$

proving that the same clustering $\mathcal{C}$ minimizes both LambdaCC and the Hamiltonian (2). In addition to this connection, we prove that *standard* LambdaCC is related to both sparsest cut and cluster deletion:

THEOREM 1 *Let $\mathcal{C}_\lambda$ be the optimal standard* LambdaCC *clustering for a fixed λ, $m = |E|$, and $n = |V|$. Define the edge density of a cluster to be the ratio between number of edges and number of pairs of nodes in the cluster.*

(a) For any λ, each cluster in $\mathcal{C}_\lambda$ has sparsest cut bounded above by λn and edge density bounded below by λ.

(b) There exists some λ^* such that $\mathcal{C}_{\lambda^*}$ is also the minimum sparsest cut partition.

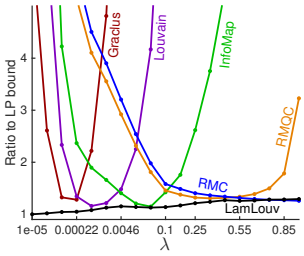
(c) For any $\lambda > m/(m+1)$, $\mathcal{C}_\lambda$ optimally solves the cluster deletion problem.

We summarize the equivalence relationships between LambdaCC and other objectives in Figure 2.

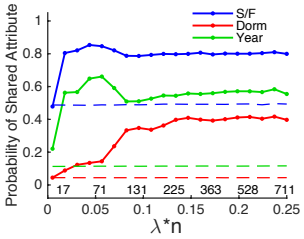
Algorithms We provide a 3-approximation for LambdaCC when $\lambda > 0.5$, and a related 2-approximation for the cluster deletion objective, using linear programming relaxations of (3). Additionally, we develop a fast heuristic method for greedily optimizing the objective which works well in practice, called Lambda-Louvain [4].

Applications and Experiments

LambdaCC has many practical applications for network analysis. In Figure 3a we illustrate that other graph clustering algorithms (e.g. Graclus or Louvain) implicitly optimize LambdaCC in different parameter regimes. The y-axis gives the ratio between a clustering's objective score and a lower bound on the optimal objective determined by solving an LP relaxation. Our method Lambda-Louvain interpolates among all these strategies when we run it for different λ, indicating that our framework can serve as a good proxy for any application in which these other methods are known to perform well.



(a) ca-GrQc network



(b) Yale Facebook Subgraph

Figure 3: (a): Other algorithms implicitly optimize LambdaCC. (b): Varying λ shows correlation between metadata attributes and communities on a university campus.

We can also use our framework to highlight the correlation between metadata attributes and communities in a social network. We illustrate this by clustering the Yale University Facebook network for various λ using Lambda-Louvain. Curves in Figure 3b trace out the probability that two people sharing the same cluster (for a fixed λ) also share the same value for a metadata attribute such as student/faculty status (blue), dorm (red), or graduation year (green). These probabilities are contrasted with the null probability of sharing a related *spurious* attribute (shown with dashed lines of the same color). This experiment indicates the extent to which each metadata attribute correlates with communities, and what value of λ displays this correlation most strongly.

References

[1] N. Bansal, A. Blum, and S. Chawla. Correlation clustering. *Machine Learning*, 56:89–113, 2004.

[2] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(026113), 2004.

[3] J. Reichardt and S. Bornholdt. Statistical mechanics of community detection. *Physical Review E*, 74(016110), 2006.

[4] N. Veldt, D. F. Gleich, and A. Wirth. A correlation clustering framework for community detection. In *(To appear) Proceedings of the 27th International World Wide Web Conference*, WWW '18, preprint: https://arxiv.org/abs/1712.05825, 2018.

CONDITIONAL TESTS FOR LOG-LINEAR ERGMS

Elizabeth Gross, Sonja Petrović, Despina Stasi

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Exponential random graph models (ERGMs) are popular choices of statistical models when using a model-based approach to network analysis. Focusing on log-linear ERGMs, we describe new algorithms for assessing the quality of fit of these models and illustrate the method on two biological datasets.

Exponential random graph models

Exponential random graph models (ERGMs) are exponential families of distributions. An ERGM is specified by its sufficient statistic T, a map from the space of all simple graphs on n vertices $\mathcal{G}_n$ to $\mathbb{R}^d$, where each of the d entries of $T(G)$ is a network statistic. A network statistic may be the number of edges in G, the number of reciprocated edges in G, the degree of a specific node in G, etc. For example, if it is decided that the network statistics of interest are the in-degree and out-degree of each vertex, then $T(G)$ is the concatenated in-degree sequence and out-degree sequence of G. Once T is specified, the ERGM defined by T is the set of all probability distributions on $\mathcal{G}_n$ given by

$$P(X = G) = \frac{1}{Z(\theta)} e^{T(G)\cdot\theta}$$

where $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ a real parameter vector, with one parameter for each network statistic, and $Z(\theta)$, the normalizing constant.

Model fitting

When fitting an ERGM to network data, first the parameters θ are estimated and then the goodness-of-fit is assessed. In goodness-of-fit testing, we are asking, *How well does the model fit the data?*.

Part of the difficulty in studying networks within a model-based setting is that statistical theory regarding fitting random graph models is still in its infancy, since it poses several challenging combinatorial and algorithmic problems. Even for the simplest network models, quantitative methods for goodness-of-fit testing are generally lacking in the statistics literature, with existing methods mainly heuristic [11], [2]. An exception are ERGMs that can be interpreted as log-linear models on contingency tables, i.e. if T is a linear function on the entries of a reasonable table representation of a graph in $\mathcal{G}_n$.

In the case where the ERGM belongs to the class of log-linear models, the goodness-of-fit question has been theoretically answered in the field of algebraic statistics [5], even if methods for effective implementation have remained elusive (see also [13] for survey). The method in [5] uses Markov bases and a Monte Carlo Markov chain algorithm to obtain a finite-sample approximation of an exact conditional test, in particular, Fisher's exact test. While the method of Markov bases is theoretically sound, it exhibits difficulties in implementation and practice as pointed out by [16]. One of the main challenges is the computational complexity of obtaining a Markov basis at the outset; a Markov basis is a set of moves that guarantees the Markov chain will be irreducible. However, since the publication of [16], there has been work to address this issue by generating Markov moves on-the-fly.

Indeed, dynamic approaches to generating Markov moves have recently been explored to remedy the computational strain of computing full bases [6], [9], [12]. For example, previous work of the authors gives a combinatorially based method for goodness-of-fit testing for the p_1 random graph model with edge-dependent reciprocation [9] and is the starting point for this study.

Approximating conditional tests

The exact conditional test implemented in [9] relies on the Markov bases methodology, but avoids the usual computational bottleneck by constructing Markov moves based on the current state of the chain, rather than computing the full basis up front. In this current work, we significantly extend the approach of [9], describing and implementing goodness-of-fit algorithms for several models beyond the p_1-model with edge-dependent reciprocation. While the methodology applies to any log-linear ERGM, we specifically focus on the β-model (see [3]), the p_1-model with three different reciprocation effects [10, 8], block versions

of these (see [15] for the β-blockmodel and [7] for the first p_1-blockmodel variants), as well as versions with structural zeros [1, Section 5.1]. We further generalize the p_1-blockmodels from [7]. These models can be viewed as the directed version of the degree-corrected stochastic block model with additional possible reciprocation effects.

To extend the MCMC algorithm, we rely on the underlying model combinatorics. In a nutshell, to construct moves, we use the fact that the sufficient statistic T is a linear map. Any linear map can be represented by a matrix, and in the models considered above, the matrix is in fact a 0–1 matrix. This 0–1 matrix naturally encodes a combinatorial object called the model hypergraph $\mathcal{H}$. In this setting, any graph G corresponds to a multiset of edges of $\mathcal{H}$, and for each move, the goal is to find another multiset of edges of having the same degree sequence. We thus design the sampling algorithms combinatorially by creating moves that preserve these degree sequences and only select allowable edges from the model hypergraph. In general, this is a very difficult sampling problem; however, for the specific models in this study, the structure of the model hypergraph allows us to break the sampling algorithm into pieces that work on subgraphs.

Data analysis of biological networks

Our method is illustrated on two experimental datasets, a neuronal network [14] and a protein-protein interaction network [4]. The neuronal network has 279 vertices and $8,693$ edges, while the protein-protein interaction network is the combination of two overlapping networks with a total of $4,344$ vertices and $9,449$ edges. Both protein-protein interaction networks and neuronal networks have been used as examples of scale-free networks, that is, networks whose degree distribution follows a power law, suggesting degree-based edge formation mechanisms in these data. Although descriptive statistics of these networks have been studied, somewhat surprisingly, there have been few studies rigorously analyzing these networks using a model-fitting approach. Our experimental results show that none of the variants of the p_1 fit the directed part of the neuronal network, implying that the edge formation in these networks is not mainly driven by the attractiveness and expansiveness of nodes. However, preliminary results for the protein-protein interaction network, suggest that the p_1 model with structural zeros fits the data well.

References

[1] Y. M. Bishop and S. E. Fienberg. Discrete multivariate analysis theory and practice. 2007.

[2] N. B. Carnegie, P. N. Krivitsky, D. R. Hunter, and S. M. Goodreau. An approximation method for improving dynamic network model fitting. *Journal of Computational and Graphical Statistics*, 24(2):502–519, 2015.

[3] S. Chatterjee, P. Diaconis, and A. Sly. Random graphs with a given degree sequence. *Annals of Applied Probability*, 21(4):1400–1435, 2011.

[4] A. I. M. Consortium. Evidence for network evolution in an arabidopsis interactome map. *Science*, 333(6042):601–607, 2011.

[5] P. Diaconis and B. Sturmfels. Algebraic algorithms for sampling from conditional distribution. *Annals of Statistics*, 26(1):363–397, 1998.

[6] A. Dobra. Dynamic Markov bases. *Journal of Computational and Graphical Statistics*, pages 496–517, 2012.

[7] S. E. Fienberg, M. M. Meyer, and S. S. Wasserman. Statistical analysis of multiple sociometric relations. *Journal of the American Statistical Association*, 80(389):51–67, 1985.

[8] S. E. Fienberg and S. S. Wasserman. Discussion of Holland, P. W. and Leinhardt, S. "An exponential family of probability distributions for directed graphs". *Journal of the American Statistical Association*, 76:54–57, 1981.

[9] E. Gross, S. Petrović, and D. Stasi. Goodness of fit for log-linear network models: Dynamic Markov bases using hypergraphs. *Annals of the Institute of Statistical Mathematics*, 2016. DOI: 10.1007/s10463-016-0560-2.

[10] P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs (with discussion). *Journal of the American Statistical Association*, 76(373):33–65, 1981.

[11] D. R. Hunter, S. M. Goodreau, and M. S. Handcock. Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258, 2008.

[12] M. Ogawa, H. Hara, and A. Takemura. Graver basis for an undirected graph and its application to testing the beta model of random graphs. *Annals of Institute of Statistical Mathematics*, 65(1):191–212, February 2013.

[13] S. Petrović. *A survey of discrete methods in (algebraic) statistics for networks*, volume 685 of *Contemporary Mathematics*, pages 260–281. American Mathematical Society, 2017.

[14] L. R. Varshney, B. L. Chen, E. Paniagua, D. H. Hall, and D. B. Chklovskii. Structural properties of the caenorhabditis elegans neuronal network. *PLoS Computational Biology*, 7(2), 2011.

[15] X. Yang. *Social Network Modeling and the Evaluation of Structural Similarity for Community Detection*. PhD thesis, Carnegie Mellon University, 2015.

[16] J. Zhang and Y. Chen. Sampling for conditional inference on network data. *Journal of the American Statistical Association*, 108(403), 2013.

TOWARD A SAMPLING THEORY FOR STATISTICAL NETWORK ANALYSIS

Harry Crane and Jianxiang Gao

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Many classical network models (e.g., stochastic blockmodels, graphons, exponential random graph models) are ill-suited for modern applications because they implicitly assume that the data is obtained by an unrealistic sampling scheme, such as vertex selection or simple random vertex sampling. More recent approaches (completely random measures and edge exchangeable models) improve somewhat upon these limitations, but leave plenty of room for further exploration of the role played by sampling in network analysis. We present here a framework that is intended to overcome theoretical and practical issues arising from the use of ill-specified network models. Within this framework we discuss how to incorporate the sampling scheme into statistical models in a way that is both flexible and insightful for modern network science applications. This talk builds on content from the recently released book *Probabilistic Foundations of Statistical Network Analysis*.

Introduction

Probabilistic models are commonly used to explain widespread empirical phenomena observed in real-world networks, such as scale-free structure and the small-world property. The preferential attachment [2] and small-world [14] models provide elegant and mathematically tractable tools for analyzing such behaviors analytically. And while some recent empirical investigations question the 'universality' of the scale-free property [3], there remains both practical and theoretical interest in understanding such behaviors, especially among statisticians [4, 7, 13].

A major outstanding question in network analysis, e.g., [6, 10, 12, 15], is the extent to which the mode of sampling affects observed network properties. For example, it is well-known that observing the subgraph obtained by simple random vertex sampling from a sparse graph produces an empty graph with high probability. Furthermore, networks obtained by degree-biased sampling from a homogeneous (i.e., not 'scale-free') graph tend to exhibit scale-free behavior, e.g., [15]. Both of these examples raise the question of how sampling scheme can be accounted for appropriately in statistical network models. To date, the role of sampling in network analysis remains poorly understood, and we shall present here some recent work directed toward gaining a better understanding.

Problem Setup

Throughout this talk, we assume a population network $\mathbf{Y}$ of size $N \geq 1$ (possibly infinite) is represented as a $\{0,1\}$-valued array $(Y_{ij})_{1\leq i,j\leq N}$, and the observed network data $\mathbf{Y}_n$ of size $1 \leq n \leq N$ is obtained by sampling from $\mathbf{Y}$. We write $\mathbf{Y}_n = \Sigma_n \mathbf{Y}$ to indicate that $\mathbf{Y}_n$ has been sampled from $\mathbf{Y}$ by some (possibly random) sampling mechanism Σ_n. The object of this talk will be to formalize the concept of random sampling by Σ_n and to explore its effects on network analysis.

In the setting of [5, Chapter 5], we specify a model for network data of all finite sample sizes $(\mathbf{Y}_n)_{1\leq n\leq N}$ as a set of finite sample models $\{\mathcal{M}_n\}_{1\leq n\leq N}$, where each $\mathcal{M}_n$ is a set of candidate distributions for $\mathbf{Y}_n$, together with a system of (random) sampling mechanisms $\{\Sigma_n\}_{1\leq n\leq N}$, where each Σ_n is a random sampling map for observing a network in $\{0,1\}^{n\times n}$ from one in $\{0,1\}^{N\times N}$. The model $(\{\mathcal{M}_n\}_{1\leq n\leq N}, \{\Sigma_n\}_{1\leq n\leq N})$ describes the two primary components of a statistical model:

- The *descriptive* component $\{\mathcal{M}_n\}_{1\leq n\leq N}$ contains the family of candidate distributions describing the variability/uncertainty in the observed network.
- The *inferential* component $\{\Sigma_n\}_{1\leq n\leq N}$ provides the context under which the observed network is to be interpreted by relating the observation $\mathbf{Y}_n$ to the population $\mathbf{Y}$, i.e., $\mathbf{Y}_n = \Sigma_n \mathbf{Y}$.

For a simple example, we define the *projective Erdős–Rényi model* by

$$\mathcal{M}_n = \{\Pr(\mathbf{Y}_n = \cdot\,; \theta) : 0 \leq \theta \leq 1\}$$

for

$$\Pr(\mathbf{Y}_n = \mathbf{y}; \theta) = \prod_{1\leq i\neq j\leq n} \theta^{y_{ij}}(1-\theta)^{1-y_{ij}}, \quad \mathbf{y} \in \{0,1\}^{n\times n},$$

and $\Pr(\Sigma_n = S_n) = 1$, for $S_n : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$ defined as the *selection* (or restriction) map, $\mathbf{y} \mapsto S_n\mathbf{y} = (y_{ij})_{1\le i,j\le n}$.

To formalize the above setup more generally, let $\mathcal{S}_{n,N}$ be the set of all *ψ-sampling maps* $S^{\psi}_{n,N} : \{0,1\}^{N\times N} \to \{0,1\}^{n\times n}$ defined for injection $\psi : [n] \to [N]$ by

$$\begin{aligned} S^{\psi}_{n,N} &: \{0,1\}^{N\times N} \to \{0,1\}^{n\times n} \\ \mathbf{y} &\mapsto S^{\psi}_{n,N}\mathbf{y} = (y_{\psi(i)\psi(j)})_{1\le i,j\le n}. \end{aligned} \tag{1}$$

For any probability distribution P on $\{0,1\}^{N\times N}$, $\mathbf{Y} \sim P$, and for Σ_n distributed randomly on $\mathcal{S}_{n,N}$ (which possibly depends on $\mathbf{Y}$), we define the *Σ_n-induced distribution* $\Sigma_n P$ on $\{0,1\}^{n\times n}$ by

$$\begin{aligned} &\Sigma_n P(\mathbf{Y}_n = \mathbf{y}) = \Pr(\Sigma_n\mathbf{Y} = \mathbf{y}) \\ &= \sum_{\psi:[n]\to[N]} \Pr(\Sigma_n\mathbf{Y} = \mathbf{y} \mid \Sigma_n = S^{\psi}_{n,N}) \Pr(\Sigma_n = S^{\psi}_{n,N}). \end{aligned}$$

These induced distributions figure into statistical analysis by allowing inferences about the optimal candidate distribution(s) in $\mathcal{M}_n$ based on $\mathbf{Y}_n$ to be extended to inferences about the optimal distribution in $\mathcal{M}_N$ as follows. Given an observation $\mathbf{Y}_n$ on $\{0,1\}^{n\times n}$ and an inference $\hat{\mathcal{P}}_n(\mathbf{Y}_n) \subseteq \mathcal{M}_n$ for the optimal candidate distribution(s) based on $\mathbf{Y}_n$, we infer

$$\hat{\mathcal{P}}_N(\mathbf{Y}_n) = \{P \in \mathcal{M}_N : \Sigma_n P \in \hat{\mathcal{P}}_n(\mathbf{Y}_n)\} \tag{2}$$

for the optimal set of distributions for the population based on $\mathbf{Y}_n$. In words, (2) is the subset of all candidate distributions for the population network $\mathbf{Y}$ that are consistent with the inferred distributions $\hat{\mathcal{P}}_n(\mathbf{Y}_n)$ and the sampling scheme Σ_n. In [5, Chapter 5], we call the model $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\Sigma_n\}_{1\le n\le N})$ *coherent* if

$$\Sigma_n\mathcal{M}_N := \{\Sigma_n P : P \in \mathcal{M}_N\} = \mathcal{M}_n \quad \text{for all } 1 \le n \le N.$$

In words, the model is coherent if and only if the model $\mathcal{M}_n$ specified for each finite sample agrees with the model induced by the population model $\mathcal{M}_N$ through the assumed sampling scheme Σ_n. As discussed in [5, Chapter 5], coherence is a logical requirement to ensure that the inference in (2) makes sense in the context of the assumed model $(\{\mathcal{M}_n\}_{1\le n\le N}, \{\Sigma_n\}_{1\le n\le N})$. It is the goal of this talk to explore potential realistic choices of $\{\Sigma_n\}_{1\le n\le N}$ for networks applications of interest.

Example: Graphons

Graphons are a widely studied class of models studied in the theoretical statistics literature [1, 9, 11]. Given a function $\phi : [0,1] \times [0,1] \to [0,1]$, define the *graphon process directed by ϕ* as the distribution on $\mathbf{Y} = (Y_{ij})_{1\le i,j\le N}$ defined by constructing the entries $\mathbf{Y}$ conditionally independently with probabilities

$$\begin{aligned} &\Pr(\mathbf{Y}_{ij} = 1 \mid U_1, U_2, \ldots; \phi) = \phi(U_i, U_j) \quad \text{and} \\ &\Pr(\mathbf{Y}_{ij} = 0 \mid U_1, U_2, \ldots; \phi) = 1 - \phi(U_i, U_j), \end{aligned}$$

for $U_1, U_2, \ldots$ i.i.d. Uniform$[0,1]$. It is well-known, e.g., [6, 8, 13], that graphon models cannot explain the sparsity and power law degree distribution observed in many real-world networks. In the context of the preceding section, graphon models are most naturally specified for $\{\Sigma_n\}_{n\ge1}$ taken (independent of $\mathbf{Y}$) to be either *selection sampling*, i.e., $\Pr(\Sigma_n = S_n) = 1$ for all $1 \le n \le N$, or *simple random vertex sampling*, i.e., $\Pr(\Sigma_n = S^{\psi}_{n,N}) = 1/N^{\downarrow n}$ for all $\psi : [n] \to [N]$, where $N^{\downarrow n} := N(N-1)\cdots(N-n+1)$. In either case, $\mathbf{Y}$ from the graphon process with parameter ϕ implies that $\Sigma_n\mathbf{Y}$ is also from a graphon process with parameter ϕ, but the independence between the sampling scheme and the population network is unrealistic in most modern networks contexts.

Example: Edge exchangeable models

Edge exchangeable network models assign equal probability to two edge-labeled graphs that are isomorphic up to relabeling of their edges [7]. In this case, two natural sampling schemes are edge selection and simple random edge sampling, defined analogously to the previous section but with the networks indexed by their edge labels instead of vertex labels. Though still quite simple, edge sampling models improve upon vertex sampling models in that they implicitly depend on the network structure—because a network is determined by its edges, the edges cannot be sampled independently of the network. It was shown in [7] that such models are able to account for sparsity and power law degree distributions through the Hollywood model. See [7] and also [5, Chapters 9–10] for more details on edge exchangeable network models.

Conclusion

In addition we will discuss other examples and emphasize the potential for a broader theory of network science in the above proposed framework.

References

[1] D. Aldous. Representations for partially exchangeable arrays of random variables. *J. Multivariate Anal.*, 11(4):581–598, 1981.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

[3] A. Broido and A. Clauset. Scale-free networks are rare. *Accessed at https://arxiv.org/pdf/1801.03400.pdf on February 16, 2018*, 2018.

[4] F. Caron and E. Fox. Sparse graphs using exchangeable random measures. *Journal of the Royal Statistical Society, Series B*, 79(5), 2017.

[5] H. Crane. *Probabilistic Foundations of Statistical Network Analysis*. Chapman-Hall, USA, 2018.

[6] H. Crane and W. Dempsey. A framework for statistical network modeling. *Accessed at arXiv:1509.08185*, 2015.

[7] H. Crane and W. Dempsey. Edge exchangeable models for interaction networks. *Journal of the American Statistical Association*, in press, 2017.

[8] P. Diaconis and S. Janson. Graph limits and exchangeable random graphs. *Rend. Mat. Appl. (7)*, 28(1):33–61, 2008.

[9] D. Hoover. Relations on Probability Spaces and Arrays of Random Variables. Preprint, Institute for Advanced Studies, 1979.

[10] S. H. Lee, P. Kim, and H. Jeong. Statistical properties of sampled networks. *Physical Review E*, 73:016102, 2006.

[11] L. Lovász and B. Szegedy. Limits of dense graph sequences. *Journal of Combinatorial Theory, Series B*, 96:933–957, 2006.

[12] P. Orbanz. Subsampling large graphs and invariance in networks. *arXiv:1710.04217*, 2017.

[13] P. Orbanz and D. Roy. Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):437–461, 2015.

[14] D. Watts and S. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.

[15] W. Willinger, D. Alderson, and J. Doyle. Mathematics and the Internet: A source of enormous confusion and great potential. *Notices Amer. Math. Soc.*, 56(5):586–599, 2009.

INFERENCE IN REGRESSION WITH NETWORK RESPONSE

Frank Marrs

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Linear regression is a common model for the influence of vectors of observable covariates $\mathbf{x}_{ij}$ on a response, as captured by $\boldsymbol{\beta}$:

$$y_{ij} = \mathbf{x}_{ij}^T \boldsymbol{\beta} + \xi_{ij}. \tag{1}$$

However, when the response y_{ij} corresponds to an edge in a network from i to j, there is inherent statistical dependence of y_{ij} and y_{ik}, for example, as both edges concern actor i. This dependence must be accounted for, as overlooking it may result in poor estimation of $\boldsymbol{\beta}$ and of its standard error. We propose and evaluate a new class of estimators for $\boldsymbol{\beta}$ and its standard error based on the assumption of exchangeability of the error network. We show that our estimator is theoretically more accurate than the current state-of-the-art estimator and demonstrate improvements in inference through simulation. Forthcoming is an implementation of our methods in an `R` package for ease of use. A preprint of an article relevant to this talk may be found on `arXiv` [4].

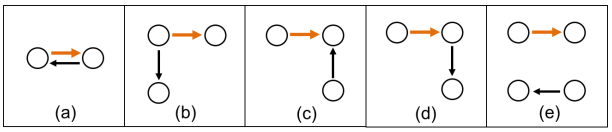


Figure 1: Five distinguishable covariance configurations between relation pairs involving the bold orange relation in an exchangeable network model: (a) reciprocal relations; (b) relations share common sender; (c) relations share common receiver; (d) shared actor is the sender of one relation and receiver of the other; (e) no shared actors among the two relations.

Background

Network data is increasingly common in pressing applications, for example, as in international commerce and adolescent development [8, 7]. One question researchers ask about such data is: how does a vector of observable covariates $\mathbf{x}_{ij}$, where i and j are indices representing nodes in the network, impact a network response y_{ij}? In this setting, the primary goal is to infer $\boldsymbol{\beta}$, including a point estimate and its standard error. When using the linear regression in (1) as a model, canonical tools for estimating $\boldsymbol{\beta}$ and its standard error may be faulty. Letting X be an $n(n-1) \times p$ matrix of covariate vectors $\{\mathbf{x}_{ij}\}_{i,j=1}^{n}$ and letting Y and $\boldsymbol{\xi}$ be the $n(n-1)$-length vectors of network responses and errors, respectively, the ordinary least squares (OLS) estimator for $\boldsymbol{\beta}$ and its variance are

$$\widehat{\boldsymbol{\beta}} = \left(X^T X\right)^{-1} X^T Y, \tag{2}$$

$$V[\widehat{\boldsymbol{\beta}}] = \left(X^T X\right)^{-1} X^T V[\boldsymbol{\xi}] X (X^T X)^{-1}. \tag{3}$$

To accomplish inference, we must estimate $V[\widehat{\boldsymbol{\beta}}]$, which amounts to estimating the error variance-covariance matrix of the errors $V[\boldsymbol{\xi}]$. Additionally, if $V[\boldsymbol{\xi}]$ is known (or well-estimated), it can be shown that the maximum likelihood estimator of $\boldsymbol{\beta}$ under normally distributed errors is $\widehat{\boldsymbol{\beta}}_{MLE} = (X^T V[\boldsymbol{\xi}]^{-1} X)^{-1} X^T V[\boldsymbol{\xi}]^{-1} Y$. In response to these two motivations, we propose an estimator of $V[\boldsymbol{\xi}]$ based on the assumption that the error network is exchangeable. This assumption is common in network models in the statistics literature, but not previously used to account dependence in a regression setting.

Existing approaches

Existing methods for accounting for network statistical dependence in regression include the application of exponential random graph models (ERGMs) [6], latent variable models [3], and post-estimation "sandwich" estimators for standard errors [2]. ERGMs provide one way to model network statistical dependence, but do not allow for generalization of the results to a population outside the sample data [5]. Additionally, ERGM specification requires modeling decisions that must be made with care and which may differ between users analyzing the same datasets. The latent variable approach requires modeling of the error network dependence directly with latent space or latent class models, for example. These models often account for network statistical dependence using random effects models, which may be difficult to estimate and force the user

to account for network dependence in the estimation of coefficients. As with ERGMs, the latent variable approach requires modeling decisions which may differ among users. The third approach accounts for network statistical dependence only after estimating coefficients $\boldsymbol{\beta}$ in (1). This approach estimates the standard errors of the coefficients using heterogenous sandwich estimators for standard errors, first proposed by [2] and termed "dyadic clustering" by [1]. These estimators are highly heterogenous, estimating each nonzero entry in the covariance matrix of the error network with a single product of residuals from the regression. Additionally, no generative model is apparent for the third approach. Our approach bridges the second and third approaches, balancing parsimony and accuracy, by assuming that the network dependence structure is jointly exchangeable after accounting for covariate information.

Exchangeable networks

Intuitively, the definition of joint exchangeability means that the node labeling has no impact on the probability distribution of the network. That is, for any permutation $\pi(.)$ on the node labels and for any two nodes i and j, the distribution of the errors $\mathbb{P}(\xi_{ij}) = \mathbb{P}(\xi_{\pi(i)\pi(j)})$. Both latent space and basic stochastic blockmodels are examples of network models that are jointly exchangeable. We show that, under the joint exchangeability assumption, the true covariance matrix for weighted network data (which we term the "exchangeable covariance matrix") has at most six unique terms: one variance and five covariances. See Figure 1 for a depiction of the five possible pairwise relations between edges in *any* jointly exchangeable network.

Results

Confidence intervals based on standard errors for $\widehat{\boldsymbol{\beta}}$ typically rely on the assumption that this estimator is asymptotically normal. Thus, we begin by proving that, when the errors are jointly exchangeable, the ordinary least squares estimator of $\boldsymbol{\beta}$ is asymptotically normal. We then propose an estimator of the six parameters of the covariance matrix for weighted network data; this estimator leads to estimates that are parsimonious and accurate. In our approach, we can choose to account for the inherent dependence – dependence that is implied by *any* jointly exchangeable network model – at the estimation stage and/or the inference stage. We prove that the standard error estimates based on our estimator are (a) asymptotically correct and (b) have lower mean-square error than the mean-square error of those based on the dyadic clustering estimator. We also demonstrate the improvement in inference when using our estimator in a simulation study, showing that our exchangeable estimator is more accurate in estimation of standard errors than the existing dyadic clustering estimator when analyzing network data; for example, the coverages in Figure 2 are closer to nominal and less variable than the dyadic clustering approach.

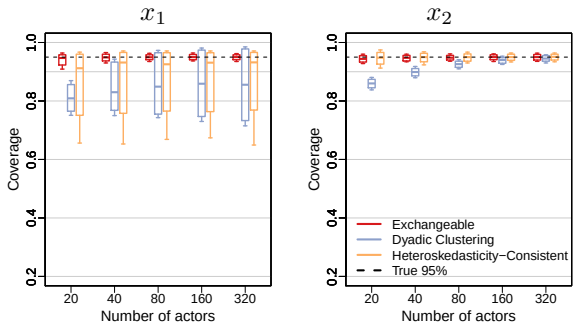


Figure 2: Probability β is in 95% confidence interval, estimated by OLS with confidence intervals estimated with various approachs, for simulation from linear model in (1) with network errors. Boxplots are across 500 random design matrices, $X = \{[x_1, x_2]_{ij}\}_{i,j=1}^{n}$.

References

[1] P. M. Aronow, C. Samii, and V. A. Assenova. Cluster–robust variance estimation for dyadic data. *Political Analysis*, 23(4):564–577, 2015.

[2] M. Fafchamps and F. Gubert. The formation of risk sharing networks. *Journal of Development Economics*, 83(2):326–350, 2007.

[3] P. D. Hoff. Multiplicative latent factor models for description and prediction of social networks. *Computational and Mathematical Organization Theory*, 15(4):261, 2009.

[4] F. W. Marrs, T. H. McCormick, and B. K. Fosdick. Standard errors for regression on relational data with exchangeable errors. *arXiv preprint arXiv:1701.05530*, 2017.

[5] C. R. Shalizi and A. Rinaldo. Consistency under sampling of exponential random graph models. *Annals of Statistics*, 41(2):508, 2013.

[6] T. A. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock. New specifications for exponential random graph models. *Sociological Methodology*, 36(1):99–153, 2006.

[7] M. Vogel, C. E. Rees, T. McCuddy, and D. C. Carson. The highs that bind: school context, social status and marijuana use. *Journal of Youth and Adolescence*, 44(5):1153–1164, 2015.

[8] A. H. Westveld and P. D. Hoff. A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict. *The Annals of Applied Statistics*, pages 843–872, 2011.

INFLUENCE ESTIMATION ON SOCIAL MEDIA NETWORKS USING CAUSAL INFERENCE

Edward K. Kao, Steven T. Smith, Danelle C. Shah, Olga Simek, MIT Lincoln Laboratory
Donald B. Rubin, Harvard University

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Estimating influence on social media networks is an important practical and theoretical problem, especially because this new medium is widely exploited as a platform for disinformation and propaganda. This paper introduces a novel approach to influence estimation on social media networks and applies it to the real-world problem of characterizing active influence operations on Twitter during the 2017 French presidential elections. The new influence estimation approach attributes impact by accounting for narrative propagation over the network using a network causal inference framework applied to data arising from graph sampling and filtering. The ability to infer high causal influence, not noticeable with activity and topological statistics, is demonstrated on real-world social media accounts that are later independently confirmed to be either directly affiliated or correlated with foreign influence operations using evidence supplied by the U.S. Congress and journalistic reports.

Introduction

The explosion of social media world-wide has created a potent medium and technology for disinformation and propaganda. Detecting and estimating influence on social media networks is the problem of inferring the impact of an account on the rest of the network, with applications in marketing on social media, influence maximization, information diffusion, and the spread of both information and disinformation in social networks.

This abstract introduces a novel approach to influence estimation on social media networks and applies it to characterizing online influence operations. This approach uses targeted collection and narrative detection to construct the narrative network (e.g. the #MacronLeaks network in Fig. 1) on which the causal influence of each vertex (i.e. account) is estimated. This abstract focuses on the causal influence estimation step. A description of the overall approach is in our paper [4].

*This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

Causal influence estimation methodology

The causal inference approach quantifies influence by accounting for causal narrative propagation over the entire network. It also accounts for several potential confounders (e.g. community membership, popularity) and removes their effects from the causal estimation. This approach is based on the network potential outcome framework [2], itself based upon Rubin's causal framework [1].

Let $G = (V,E)$ be a graph with N vertices $V = \{v_1, v_2, \ldots, v_N\}$, whose edges are denoted by the observed interactions between v_i and v_j, let $\mathbf{A} = (a_{ij})$ be an N-by-N random matrix of social influence of v_i on v_j with Poisson rate determined by the graph data G, and $\boldsymbol{Z}$ be a binary N-vector of narrative sources (a.k.a. treatment vector). The fundamental quantity is the *network potential outcome* of each vertex, denoted $Y_i(\boldsymbol{Z}, \mathbf{A})$, under exposure to the narrative through the source vector $\boldsymbol{Z}$ and influence network $\mathbf{A}$. In the analysis below, vertices are user accounts, edges are retweets of a specific narrative, and the potential outcomes are the number of tweets in the narrative. The impact ζ_i of each vertex on the narrative is defined using the network potential outcomes,

$$\zeta_i(\boldsymbol{z}) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} \big(Y_j(\boldsymbol{Z} = \boldsymbol{z}_{i+}, \mathbf{A}) - Y_j(\boldsymbol{Z} = \boldsymbol{z}_{i-}, \mathbf{A}) \big). \qquad (1)$$

This causal estimand is the average difference between the individual outcomes with v_i as a source s.t. $\boldsymbol{z}_{i+} = (z_1, \ldots, z_i := 1, \ldots, z_N)^{\mathrm{T}}$, versus v_i *not* a source s.t. $\boldsymbol{z}_{i-} = (z_1, \ldots, z_i := 0, \ldots, z_N)^{\mathrm{T}}$. This impact is the average (per vertex) number of additional tweets generated by an user's participation. The source is *uniquely impactful* if it is the only source.

It is impossible to observe the outcomes at each vertex with both exposure conditions under source vectors $\boldsymbol{z}_{i+}$ and $\boldsymbol{z}_{i-}$; therefore, the missing potential outcomes must be estimated. This can be accomplished by modeling the potential outcomes. After estimating the model parameters on the observed outcomes and vertex covariates, missing potential outcomes in the causal estimand ζ_i can be imputed using the fitted model. In this analysis, potential outcomes are modeled using a Poisson generalized linear model (GLM) with the canonical log

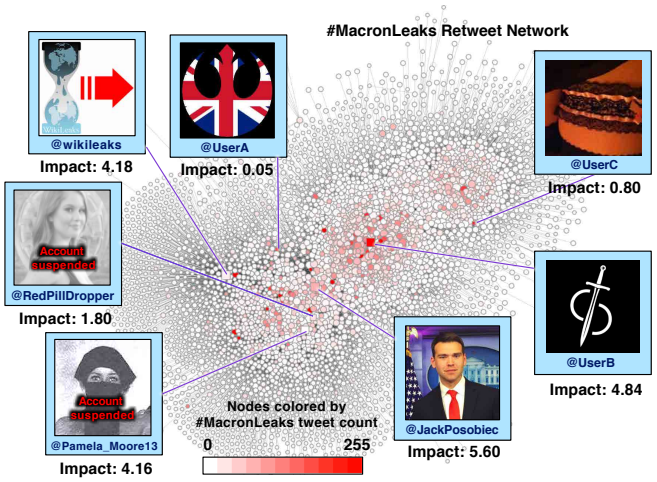


Figure 1: Influence estimation on the #MacronLeaks narrative network. Vertices are accounts and edges are retweets. Vertices are colored by the number of #MacronLeaks tweets, and sized by its out-degree. Impact is the average number of additional tweets generated by an user's participation [Eq. (1)].

link function and linear predictor coefficients $(\tau, \gamma, \boldsymbol{\beta}, \mu)$ corresponding to the source indicator Z_i, n-hop exposure vector $s_i^{(n)}$, the covariate vector $\boldsymbol{x}_i$, and the baseline outcome. The covariate vector $\boldsymbol{x}_i$ includes the potential social confounders such as popularity and community membership. The GLM model for the potential outcomes is,

$$\log \lambda_i = \tau Z_i + \left(\sum_{n=1}^{N_{\text{hop}}} \prod_{k=1}^{n} \tau \gamma_k s_i^{(n)} \right) + \boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i + \mu + \varepsilon_i, \qquad (2)$$

with $Y_i(\boldsymbol{Z}, \mathbf{A}) \sim \text{Poisson}(\lambda_i)$. The first term in the linear predictor τZ_i represents the primary effect on the source. The second term $\sum_{n=1}^{N_{\text{hop}}} \prod_{k=1}^{n} \tau \gamma_k s_i^{(n)}$ represents the accumulative social influence effect from n-hop exposures $s_i^{(n)}$ to the source, where each coefficient γ_k captures the decay of the effect over each additional hop. The third term $\boldsymbol{\beta}^{\mathrm{T}} \boldsymbol{x}_i$ is the effect of the unit covariates $\boldsymbol{x}_i$ including the potential social confounders such as popularity and community membership. The fourth term, μ, is the baseline effect on the entire population. The last term $\varepsilon_i \sim \text{Normal}(0, \sigma_\varepsilon^2)$ gives independent and identically distributed variation for heterogeneity between the units. The amounts of social exposure at the nth hop are determined by $(\mathbf{A}^{\mathrm{T}})^n \boldsymbol{Z}$. This captures all exposure to the sources propagated on the influence network. Lastly, to model the diminishing return of additional exposures, the (nonnegative) log-exposure is used, $\boldsymbol{s}^{(n)} = \log\left((\mathbf{A}^{\mathrm{T}})^n \boldsymbol{Z} + 1\right)$. Joint inference of the model parameters τ, γ, $\boldsymbol{\beta}$, μ, and the influence matrix $\mathbf{A}$ is done through Monte Carlo Markov Chain and Bayesian regression.

Table 1: #MacronLeaks narrative network screen names, tweets (T), total retweets (TRT), followers (F), initial times (on 5 May), PageRank centrality (PR), and estimated influence.

Screen name	T	TRT	F	1st time	PR	**Impact***
@JackPosobiec	95	47 k	261 k	18:49	2.84	**5.60**
@RedPillDropper	32	8 k	8 k	19:33	2.86	**1.80**
@UserA†	256	59 k	1 k	19:34	27.08	**0.05**
@UserB†	260	54 k	3 k	20:25	57.05	**4.84**
@wikileaks	25	63 k	5515 k	20:32	2.80	**4.18**
@Pamela_Moore13	4	4 k	54 k	21:14	2.79	**4.16**
@UserC†	1305	51 k	$<$ 1 k	22:16	6.36	**0.80**

*Influence estimate from Eq. (1) applied to data
†Anonymized screen names of currently active accounts

Results

The proposed approach identifies highly influential accounts that are independently confirmed by the U.S. Congress and journalistic reports, on the #MacronLeaks narrative network during the 2017 French presidential elections, shown in Fig. 1 and Table 1. For example, @wikileaks and @JackPosobiec have been reported to be influential instigators of the #MacronLeaks narrative [3]. A new finding is the high impact of an account known to be tied to foreign influence operations: @Pamela_Moore13 [5]. Other accounts that are less well-known are also estimated to have high influence, e.g. @UserB, who serves as a bridge into the predominantly French-speaking subgraph (the cluster seen in the middle of Fig. 1).

These results highlight the ability of causal influence estimation to infer high impact beyond simple activity and topological statistics. For example, the highly active accounts @UserA and @UserC tweeted about #MacronLeaks many times, but were not estimated to have high impact. In contrast, some accounts with only a few #MacronLeaks tweets and lower centrality are estimated to have high causal impact. As observed in Fig. 1, higher out-degree is correlated with impact, but provides only partial information for influence estimation. E.g. @Pamela_Moore13 has lower out-degree count than @wikileaks, @JackPosobiec, and @UserB, but nevertheless has high impact from her strong influence on the vertices near her in the retweet network.

References

[1] Imbens G.W. and Rubin D.B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences*, Cambridge University Press.

[2] Kao E.K. (2017). Causal Inference Under Network Interference: A Framework for Experiments on Social Networks, Ph.D. thesis, Harvard U.

[3] Marantz A. (7 May 2017). The far-right American nationalist who tweeted #MacronLeaks, *The New Yorker*. Accessed January 2018.

[4] Smith S.T., Kao E.K., Shah D.C., Simek O., and Rubin D.B. (2018). Influence Estimation on Social Media Networks Using Causal Inference. In *Proc. IEEE SSP Workshop*.

[5] U.S. House Permanent Select Committee on Intelligence. (1 November 2017). Exhibit of the user account handles that Twitter has identified as being tied to Russia's "Internet Research Agency." Accessed January 2018.

PAIRWISE VERSUS MULTIVARIATE CONSTRUCTIONS OF CO-OCCURRENCE NETWORKS

Jason Cory Brunson, Reinhard C. Laubenbacher

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Network representations of association data are often used to explore dependencies among variables, including logical variables in ecology and medicine. The pairwise construction cannot account for covariance structure, while multivariate modeling is infeasible in many practical settings. We use empirical and simulated data to test the dependence of common network-analytic results on the method of construction.

Motivation

Network representations of correlation data are widely used to explore variable dependencies. A special use case is the analysis of co-occurrence data, which encodes values of several logical variables for each member of a population. Population ecologists use co-occurrence networks to represent biotic interactions among species observed at different sites, and the same approach is increasingly taken to comorbidity relations among health disorders. In low-dimensional settings, the construction has matured from aggregations of pairwise associations to discretizations of the covariance structure in multivariate binomial regression models. In even moderately high-dimensional settings, covariance estimation can be computationally infeasible, and pairwise constructions dominate.

In contrast to classical networks, co-occurrence networks are generated by statistical inference rather than from observed relations. Pairwise correlations that ignore confounding by other variables are subject to biases that may impact their magnitude, sign, and discernibility, and both distance-based and motif-based properties are sensitive to link addition and deletion. Limited attention has been given to the impact of such confounding on network-analytic (graph-theoretic) properties of the resulting co-occurrence network.

Network constructions

Conventional pairwise (PW) co-occurrence networks are constructed pairwise from frequency tables $\begin{pmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{pmatrix}$ with $\mathrm{E}(a_{ij}) \propto \Pr(Y_1 = i \wedge Y_2 = j)$, for each pair of logical variables (y_1, y_2), using a statistical test to identify links and an association measure to assign them weights. We use Fisher's exact test to identify links and the tetrachoric correlation coefficient r_t [2] to quantify associations. r_t is the value of ρ that maximizes the likelihood of observed frequencies under the latent variable model

$$\begin{aligned} \Pr(Y_i) &= \Pr(Z_i > 0) \\ (Z_1, Z_2) &\sim N(\mu, \Sigma) \\ \Sigma &= \begin{pmatrix} {\sigma_1}^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & {\sigma_2}^2 \end{pmatrix} \end{aligned}$$

This choice facilitates comparisons with two other constructions: First, the assumption of normality implies that the partial correlations $r'_{ij} = -\kappa_{ij}/\sqrt{\kappa_{ii}\kappa_{jj}}$, where $\Sigma^{-1} = \mathrm{K} = (\kappa_{ij})$, are standardized regression coefficients from multiple linear regression models that treat each variable as a response to all others [1]. The resulting partial correlation (PC) network is based on inference from multiple rather than simple regression. Second, the multivariate, hierarchical joint distribution (JD) model, in which the latent variable means $\mu_i \sim N(\nu, \varsigma)$ are taken to have been sampled from a higher-order distribution [3], generalizes r_t to any number of logical response variables. The estimated covariance matrix $\hat{\Sigma}$ is normalized to a correlation matrix $\hat{\mathrm{P}}$ for comparison with r_t and r'_t.

Datasets

One set of occurrence data comes from the National Ambulatory Medical Care Survey (NAMCS), distributed annually to a nationally representative set of physicians who record patient demographic and health indicators over a two-week period. Among the indicators are 13 chronic diseases. We group the 2011 NAMCS encounters by physician and draw a 50% cluster random sample.

Of our constructions, only JD yields a generative model. We will simulate occurrence data from these models across a range of covariance structures, in order to compare the networks aggregated from simulated data against the underlying relations.

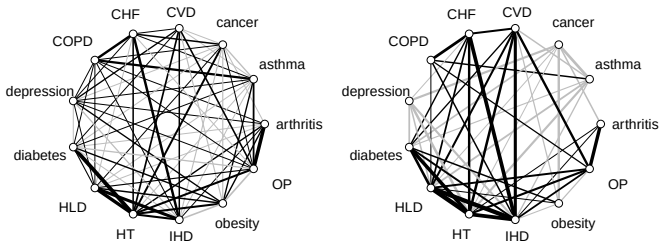


Figure 1: Co-occurrence networks for 13 chronic diseases from a common dataset using the PC (left) and JDM (right) constructions. Link thickness (color) encodes magnitude (sign) of correlation estimate.

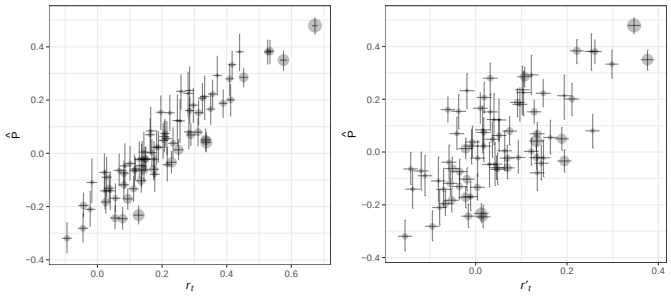


Figure 2: Maximum-likelihood estimates (posterior means) and confidence (credible) intervals for correlations between 78 pairs of chronic diseases using three models.

Sensitivity analysis

We construct co-occurrence networks using the above methods and assess the consistency of structural properties commonly used in co-occurrence network analyses. In addition to correlation estimates and their statistical evidence, we are examining several distance- and motif-based statistics, including triad closure, mean geodesic distance, and node centrality measures. We concern ourselves both with the sensitivity of their values and with how well they can be predicted in one model from another.

Results

This is a work in progress, and we report some preliminary observations here.

Even in low-dimensional settings, network dependencies on model construction can be pervasive. Fig. 1 shows two networks constructed using different methods from the NAMCS sample. The PW network (not shown) is nearly complete with only one negative link, whereas almost half of the links are negative in the PC and JD networks. Even then, the signs of the links vary significantly between the latter constructions.

Fig. 2 shows that these dissimilarities obscure a highly predictive relationship between the PW (r_t) and JD ($\hat{\mathrm{P}}$) correlations: r_t and $\hat{\mathrm{P}}$ are clearly linearly related with slope near 1 and greater coefficient of determination. The comparisons with the partial correlations ($\hat{\mathrm{P}}$ versus r'_t shown) indicate that the negative shift from r_t to $\hat{\mathrm{P}}$ (-0.21) is due to the confounding of other variables (r_t versus r'_t), but also suggests a trade-off in the partial correlations between correcting the shape of the distribution and distorting the rankings of the values.

As suggested by a reviewer, we evaluated the sensitivity of each comorbidity $r_{ij} = r'_t(i,j)$ to each disorder k as the difference $\Delta r_{ij|k} = (r_{ij|k} - r_{ij})$, calculating $r_{ij|k}$ as r_{ij} on the dataset with variable k omitted. The 6 sensitivities with greatest magnitudes are listed in Table 1 (all positive, though 70% of sensitivities were negative). We observed no relationship between $\Delta r_{ij|k}$ and r_{ij}.

Disorder	Comorbidity	r_{ij}	$\Delta r_{ij\|k}$
HT	diabetes–HLD	0.107	0.166
HLD	HT–IHD	0.221	0.123
HT	HLD–IHD	0.261	0.108
HT	diabetes–obesity	0.128	0.091
HT	diabetes–IHD	−0.061	0.086
HT	CVD–diabetes	−0.037	0.084

Table 1: Greatest sensitivities of estimated comorbid associations to removals of single disorders from the network.

To trace the dependency of graph-theoretic properties on the choice of model, we are examining a family of networks that interpolate between the PW and PT networks by converting sample to partial correlations within variable clusters. Graph properties exhibit no consistent patterns along this interpolation, except those due to changes in link density; we expect to get more useful results using higher-dimensional datasets.

References

[1] S. Epskamp and E. I. Fried. A Tutorial on Regularized Partial Correlation Networks. Dec 2017.

[2] D. Fritz. *Polychoric and Polyserial Correlations*, pages 69–74. Wiley, 1988.

[3] L. J. Pollock, R. Tingley, W. K. Morris, N. Golding, R. B. O'Hara, K. M. Parris, P. A. Vesk, and M. A. McCarthy. Understanding co-occurrence by modelling species simultaneously with a Joint Species Distribution Model (JSDM). *Methods in Ecology and Evolution*, 5(5):397–406, 2014.

LARGE DEGREE ASYMPTOTICS AND RECONSTRUCTION THRESHOLDS OF ASYMMETRIC ISING CHANNELS

WENJIAN LIU* AND NING NING†

1. INTRODUCTION

Determining the reconstruction threshold of a Markov random field in probability, as the interdisciplinary subject, has attracted more and more attention from probabilists, statistical physicists, biologists, etc. In fact, the investigation of the reconstruction problem originated from spin systems in statistical physics by establishing that the reconstruction threshold happens to be the threshold for extremality of the infinite-volume Gibbs measure with free boundary conditions [9]. It is shown that the reconstruction bound determines the efficiency of the Glauber dynamics on trees and random graphs [1, 13], for example, the mixing time undergoes a phase transition at the reconstruction threshold. The reconstruction threshold is also believed to play an important role in a variety of other contexts, such as the efficiency of reconstructing phylogenetic ancestors in evolutionary biology [7, 16], communication theory in the study of noisy computation [8], network tomography [2] (given end-to-end delays in a computer network, infer the link delays in its interior).

We start with the following broadcasting process that stands as a discrete, irreducible, aperiodic, and reversible Markov chain. Let $\mathbb{T} = (\mathbb{V}, \mathbb{E}, \rho)$ be a tree with nodes $\mathbb{V}$, edges $\mathbb{E}$ and root $\rho \in \mathbb{V}$. Each edge of the tree acts as a channel on a finite characters set $\mathcal{C}$, whose elements are configurations on $\mathbb{T}$, denoted by σ. Next set a probability transition matrix $\mathbf{M} = (M_{ij})$ as the noisy communication channel on each edge. The state of the root ρ, denoted by σ_ρ, is chosen according to an initial distribution π on $\mathcal{C}$. This symbol is then propagated in the tree as follows. For each vertex v having as a parent u, the spin at v is defined according to the probabilities

$$\mathbf{P}(\sigma_v = j \mid \sigma_u = i) = M_{ij}$$

with $i, j \in \mathcal{C}$. In this project, we will restrict our attention to regular d-ary trees, that is the infinite rooted tree where every vertex has exactly d offspring (every vertex has degree $d+1$ except the root which has degree d). Let $\sigma(n)$ denote the spins at distance n from the root and let $\sigma^i(n)$ denote $\sigma(n)$ conditioned on $\sigma_\rho = i$. The objective model taken into account is the asymmetric binary channel with the configuration set $\mathcal{C} = \{1, 2\}$, whose transition matrix is of the form

$$\mathbf{M} = \frac{1}{2}\left[\begin{pmatrix} 1+\theta & 1-\theta \\ 1-\theta & 1+\theta \end{pmatrix} + \Delta \begin{pmatrix} -1 & 1 \\ -1 & 1 \end{pmatrix}\right],$$

where Δ is used to describe the deviation of $\mathbf{M}$ from the symmetric channel and obviously there is a restriction of $|\theta| + |\Delta| \leq 1$. (Figure (a))

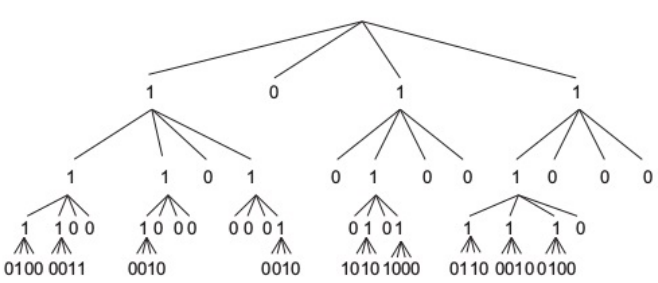
The problem of reconstruction is the following: consider all the symbols received at the vertices of the n^{th} generation. Does this configuration contain a non-vanishing information on the letter transmitted by the root, as n goes to ∞? (Figure (b))

Definition 1 *The reconstruction problem for the infinite tree $\mathbb{T}$ is* **solvable** *if for some $i, j \in \mathcal{C}$,*

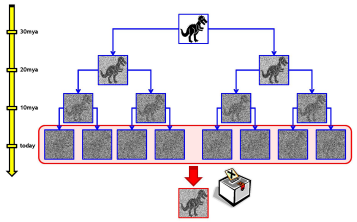
$$\limsup_{n\to\infty} d_{TV}(\sigma^i(n), \sigma^j(n)) > 0$$

where d_{TV} is the total variation distance. When the $\limsup$ *is* 0 *we will say the model has* **non-reconstruction** *on $\mathbb{T}$.*

* Department of Mathematics and Computer Science, Queensborough Community College, City University of New York
Email: wjliu@qcc.cuny.edu
† Department of Statistics and Applied Probability, University of California, Santa Barbara
Email: ning@pstat.ucsb.edu.



(a) d-ary tree



(b) Reconstruction

2. Main Results

For the binary symmetric channel, it was shown in [4] that the reconstruction problem is solvable if and only if $d\theta^2 > 1$, which we refer to as the Kesten-Stigum bound [10, 11]. For all other channels, it was also known and easy to prove that $d\theta^2 > 1$ implies solvability, while proving non-reconstructibility turned out to be harder. Although coupling arguments easily yield nonreconstruction, these arguments are typically not tight. A natural approach to non-reconstructibility is to analyze recursions in terms of random variables each of whose values is the expectation of the chain at a vertex, given the state at the leaves of the subtree below it, and corresponding probabilities. Mossel [15, 17] showed that the Kesten-Stigum bound is not the bound for reconstruction in the binary-asymmetric model with sufficiently large asymmetry or in the Potts model with sufficiently many characters, shed the light on exploring the tightness of the Kesten-Stigum bound.

Furthermore, Proposition 12 in [15] implies that for any asymmetric channel, given d and π, reconstruction is monotone in $|\theta|$, say, there exist the thresholds $\theta^- < 0 < \theta^+$ such that there is non-reconstruction when $\theta \in (\theta^-, \theta^+)$ while it is reconstructible given $\theta < \theta^-$ or $\theta > \theta^+$. But exact thresholds for non-solvability had not been known until [5], in which Borgs et al displayed a delicate analysis of the moment recursion on a weighted version of the magnetization, and thus achieved a breakthrough result. However this theorem has just established the existence of Δ_0 without estimating the range to keep Kesten-Stigum bound tight. Inspired by the refined recursive equations of vector-valued distributions and concentration analyses, which had been successfully engaged in [18, 12] to confirm much of the picture predicted by Mézard and Montanari [14], we are able to present the critical relationship between Δ and θ to preserve tightness of the Kesten-Stigum bound.

Theorem 2.1 *When $\Delta^2 > (1-\theta)^2/3$, for every d the Kesten-Stigum bound is not tight. Moreover, there exists an asymptotic result of the reconstruction threshold for fixed π and d goes to infinity:*

$$\lim_{d\to\infty} d\Theta^2 = C_\pi$$

where $0 < C_\pi < 1$ is a constant depending only on π.

Furthermore with the assistance of the central limit theorem and gaussian approximation, we figure out the precise condition to keep the tightness of the Kesten-Stigum bound for fixed π and large d.

Theorem 2.2 *When $\Delta^2 < (1-\theta)^2/3$, there exists a $D = D(\pi) > 0$ such that for $d > D$ the Kesten-Stigum bound is sharp, that is*

$$\theta^+ = d^{-1/2} \quad \text{and} \quad \theta^- = -d^{-1/2}.$$

Furthermore there is non-reconstruction at the Kesten-Stigum bound, when $\theta = \theta^+$ or θ^-.

3. Proof Sketch

The ideas and techniques used to prove Theorem 2.1 and Theorem 2.2 can be seen as the following. One standard to classify reconstruction and nonreconstruction is to analyze the quantity x_n: the probability of giving a correct guess of the root given the spins $\sigma(n)$ at distance n from the root, minus the probability of guessing the root randomly which is π_1 in this case. Nonreconstruction means that the mutual information

between the root and the spins at distance n goes to 0 as n tends to infinity. It can be established that x_n is always positive and the nonreconstruction is equivalent to $\lim_{n\to\infty} x_n = 0$.

Our analysis is similar to Borgs et al. [5], Chayes et al. [6] in the context of spin–glasses, and Sly [18]. In order to research the reconstruction, according to the Markov random field property, we establish the distributional recursion and moment recursion, by analyzing the recursive relation between the nth and the $(n+1)$th generations' structure of the tree.

Furthermore, we display that the interactions between spins become very weak, if they are sufficiently far away from each other. Therefore, we can obtain a nonlinear dynamical system. If x_n is sufficiently small, we are able to develop the concentration analysis and achieve the approximation to the dynamical system:

$$x_{n+1} \approx d\theta^2 x_n + \frac{1-6\pi_1\pi_2}{\pi_1\pi_2^2}\frac{d(d-1)}{2}\theta^4 x_n^2$$

The sign of quadratic coefficient, say, $1-6\pi_1\pi_2$ will play a crucial role in the asymptotic behavior of x_n. When $\Delta^2 > (1-\theta)^2/3$, that is, $1-6\pi_1\pi_2 > 0$, if $d\theta^2$ is sufficiently close to 1, then x_n does not converge to 0 and hence there is reconstruction beyond the Kesten-Stigum bound. The second stage is to find this new reconstruction threshold other than the Kesten-Stigum bound. Firstly we will show that when degree d is large the interactions between spins become very weak. Then using the Central Limit Theorem, we approximate this collection of small independent interactions to show that the reconstruction function can be approached by a new Gaussian approximation function $g(s)$, that is, $x_{n+1} \approx g(d\theta^2 x_n)$. Last explore first several major terms of the Maclaurin series of $g(s)$, and thus we would be able to figure out the reconstruction threshold by discussing the fixed point of $g(s)$.

On the other side, when $\Delta^2 < (1-\theta)^2/3$, applying the large degree asymptotics again yields $g(s) < s$, and it implies $\lim_{n\to\infty} x_n = 0$, that is, there is nonreconstruciton.

REFERENCES

[1] BERGER, N., KENYON, C., MOSSEL, E. and PERES, Y.: Glauber dynamics on trees and hyperbolic graphs. *Probab. Theory Related Fields* **131** 311–340. (2005).

[2] BHAMIDI, S., RAJAGOPAL, R., ROCH, S.: Network delay inference from additive metrics.: *Random Structures Algorithms* **37** 176–203. (2010).

[3] BISCONTI, C., CORALLO, A., FORTUNATO, L., GENTILE, A. A., MASSAFRA, A. and PELL, P.: Reconstruction of a real world social network using the Potts model and Loopy Belief Propagation. *Frontiers in psychology.* 6. (2015).

[4] BLEHER, P. M., RUIZ, J. and ZAGREBNOV, V. A.: On the purity of the limiting Gibbs state for the Ising model on the Bethe lattice. *J. Statist. Phys.* **79** 473–482. (1995).

[5] BORGS, C., CHAYES, J. T., MOSSEL, E. AND ROCH, S.: The Kesten-Stigum reconstruction bound is tight for roughly symmetric binary channels. *FOCS, IEEE Comput. Soc.* 518–530. Berkeley, CA. (2006)

[6] CHAYES, J. T., CHAYES, L., SETHNA, J. P., & THOULESS, D. J.: A mean field spin glass with short-range interactions. *Communications in Mathematical Physics.* 106(1), 41-89. (1986).

[7] DASKALAKIS, C., MOSSEL, E. and ROCH, S.: Optimal phylogenetic reconstruction. In *STOC'06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing* 159–168. ACM, New York. (2006).

[8] EVANS, W., KENYON, C., PERES, Y. and SCHULMAN, L. J.: Broadcasting on trees and the Ising model. *Ann. Appl. Probab.* **10** 410–433. (2000).

[9] GEORGII, H. O.: *Gibbs Measures and Phase Transition.* de Gruyter, Berlin. (1988).

[10] KESTEN, H. and STIGUM, B. P.: Additional limit theorems for indecomposable multidimensional Galton-Watson processes. *Ann. Math. Statist.* **37** 1463–1481. (1966).

[11] KESTEN, H. and STIGUM, B. P.: Limit theorems for decomposable multi-dimesional Galton-Watson processes. *J. Math. Anal. Appl.* **17** 309–338. (1966).

[12] LIU, W. and NING, N.: The Tightness of the Kesten–Stigum Reconstruction Bound of Symmetric Model with Multiple Mutations. *Journal of Statistical Physics*, **122(1)** 1–25. (2017). https://doi.org/10.1007/s10955-017-1937-1

[13] MARTINELLI, F., SINCLAIR, A. and WEITZ, D.: Fast mixing for independent sets, colorings, and other models on trees. *Random Structures Algorithms* **31** 134–172. (2007).

[14] MÉZARD, M. and MONTANARI, A.: Reconstruction on trees and spin glass transition. *J. Stat. Phys.* **124** 1317–1350. (2006).

[15] MOSSEL, E.: Reconstruction on trees: beating the second eigenvalue. *Ann. Appl. Probab.* **11** 285–300. (2001).

[16] MOSSEL, E.: Phase transitions in phylogeny. *Trans. Amer. Math. Soc.* **356** 2379–2404. (2004).

[17] MOSSEL, E.: Survey: information flow on trees. In *Graphs, morphisms and statistical physics. DIMACS Ser. Discrete Math. Theoret. Comput. Sci.* **63** 155–170. Amer. Math. Soc., Providence, RI. (2004).

[18] SLY, A.: Reconstruction for the Potts model. *Ann. Probab.* **39** 1365–1406. (2011).

SIMPLICIAL CLOSURE AND HIGHER-ORDER LINK PREDICTION

Austin R. Benson, Rediet Abebe, Michael T. Schaub, Ali Jadbabaie, and Jon Kleinberg

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Networks provide a powerful abstraction for complex systems throughout the sciences, by representing the underlying set of pairwise interactions. But much of the structure within these systems involves interactions that take place among more than two nodes at once, and while these types of higher-order interactions are ubiquitous, there has been a lack of analytical tools for evaluating higher-order models. Here we propose a general framework for evaluating higher-order structure based on link prediction, a task in which we seek to predict future interactions from a system's structure and past history. Our prediction methodology reveals phenomena that are not captured by standard network representations, in data drawn from biology, medicine, social networks, and the Web.

Higher-order network data and models. The basic premise of network models is to represent the elements of the underlying system as nodes, and to use the links of the network to capture pairwise relationships—in this way, a social network can represent the friendships between pairs of people; a Web graph can encode links among Web pages or topic categories; and a biological network can represent the interactions among pairs of biological molecules or components. But much of the structure in these systems involves *higher-order interactions* on more than two entities at once: people often communicate or interact in social groups, not just in pairs; associative relations among ideas or topics often involve the intersection of multiple concepts; and joint protein interactions in biological networks are associated with important phenomena. These types of higher-order interactions are apparent even in the standard genres of datasets used for network analysis; for example, co-authorship networks are built from data in which larger groups write papers together; similarly, email networks are based on messages that often have multiple recipients. While higher-order structure is not captured by a graph, there are formal models for such data, including set systems [2], hypergraphs [1], simplicial complexes [3], and bipartite affiliation graphs [6].

Evaluation with higher-order link prediction. Despite the existence of formalisms for higher-order structure, it has been challenging to adapt the empirical methodology developed for graph and network data to the higher-order case, due to the lack of general frameworks for evaluating models of higher-order structure. Here we propose such a framework, drawing on the concept of *link prediction*, a cornerstone problem in network analysis [4, 5].

Link prediction is a means of evaluating network models by taking network data that evolves over time and seeing how well a given model predicts the appearance of new links—for example, new collaborations appearing in a co-authorship network, or new messages between pairs of people in an email network. Link prediction is valuable both for methodological reasons and also in concrete applications. Methodologically, asking whether one model is significantly better than another at predicting new links provides a data-driven way of assessing the effectiveness of the models. But since link prediction cuts across many disciplines, it also has a range of direct applications, including predicting friendships in social networks, inferring new relationships between genes and diseases, and suggesting novel connections in the scientific community.

Here, we introduce an analogue of link prediction for higher-order structure, providing a general framework for evaluating models in any data where this type of structure evolves over time through the appearance of new, higher-order interactions — for example, predicting which sets (rather than just pairs) of authors will write a paper together, or which sets of people will appear as joint recipients on a new email message. We study the temporal evolution of 19 network datasets from a variety of domains, including social networks, online communication, and biomedicine, where each dataset is a collection of timestamped sets of nodes, which we call *simplices*. The nodes in each given simplex take part in a shared interaction at the given timestamp (Fig. 1A). For example, in a co-authorship network, a simplex corresponds to the set of authors of a publication.

The basic premise in link prediction—whether pairwise

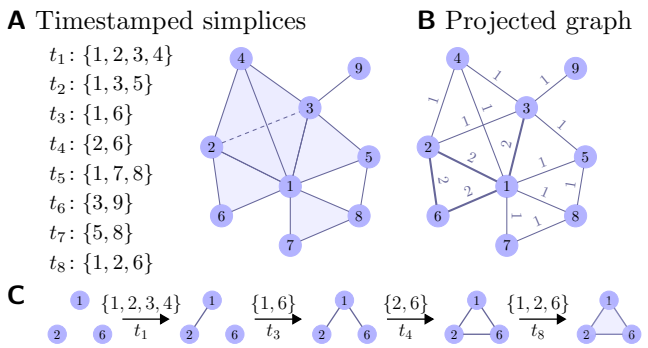


Figure 1: Higher-order network models, open and closed closed triangles, and simplicial closure. (**A**) Example dataset comprised of eight timestamped simplices and nine nodes. The dataset has seven closed triangles and one open triangle ($\{1, 5, 8\}$). (**B**) The "projected graph" of the dataset. The weight of an edge is the number of times the two end points of the edge appeared in a simplex together. (**C**) Simplicial closure of nodes 1, 2, and 6. Before closing, the three nodes induce several subgraph in the projected graph over time. For example, the nodes form an open triangle at time t_4, which persists until time t_8 when the simplex closes.

or higher-order—is to use properties of the structure up to some time t to predict the appearance of new interactions after t. To develop a set of candidate structural features, we study the *projected graph* of the system, in which two nodes are joined by an edge of weight w if they have been involved in w simplices before time t (Fig. 1B). If there are edges among all pairs from a given set of k nodes—forming a k-clique in the projected graph—this might be because (i) these k nodes were part of a single simplex together, or (ii) because each pair was part of a simplex, although all k were never part of the same simplex. In the former case, we say the k nodes form a *closed* clique, while in the latter case we say they form an *open* clique. Many of the new simplices that form in our data consist of k nodes that had previously constituted an open k-clique in the projected graph; we say that the appearance of the new simplex on these k nodes is an instance of *simplicial closure*, the conversion of an open structure to a closed one (Fig. 1C). Simplicial closure is distinct from the well-known phenomenon of *triadic closure* in social networks, since triadic closure modifies the structure of the underlying pairwise interactions, whereas simplicial closure adds a new higher-order interaction without changing the pairwise structure of the projected graph. We use these pairwise structures in the projected graph, together with higher-order structure, as the basic features in our methods for higher-order link prediction.

Results. Among other results, we find the following within our framework. First, there is enormous diversity across datasets in basic higher-order structural parameters such as the fraction of k-cliques that are open. In particular, for each of our datasets, we look at the edge density in the projected graph and at the fraction of open 3-cliques, and we see that most combinations of values are possible. Second, there is an interesting trade-off in the relative predictive power of edge density and edge weight. For the case of $k = 3$, for example, a greater number of edges among the k nodes is predictive of the arrival of a simplex; but higher weight on these edges is also predictive. Finally, the link prediction problem for higher-order structure exhibits some fundamental differences from traditional link prediction with pairwise interactions. In the traditional link prediction problem, it is valuable to use information contained in paths of non-trivial length between two nodes u and v for predicting a link between them—for example, PageRank-like measures and other forms of path enumeration are effective [4]. But we find that to predict the formation of a simplex on nodes u, v, and w, it is difficult to improve on the purely *local* information contained in the three-dimensional vector of edge weights for (u, v), (v, w), and (u, w) in the projected graph. This appears to arise from the ability of a k-tuple of nodes, for $k \geq 3$, to contain rich local information in its interactions among subsets of size $k - 1$—a phenomenon that has no natural analogue in the case of $k = 2$, and which renders the prediction problems qualitatively more distinct than one might initially suppose. Despite the power of local information, we still find that methods based on supervised learning, combining multiple structural features, can produce improvements over any one feature in isolation.

References

[1] C. Berge. *Hypergraphs.* Elsevier, 1989.

[2] P. Frankl. Extremal set systems. In *Handbook of combinatorics*, volume 1. 1995.

[3] A. Hatcher. *Algebraic topology.* Cambridge Univ. Press, 2002.

[4] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Assoc. Inf. Sci. Technol.*, 2007.

[5] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Stat. Mech. and its Applications*, 2011.

[6] M. E. J. Newman, D. J. Watts, and S. H. Strogatz. Random graph models of social networks. *PNAS*, 2002.

UNDETECTABLE CYBER-PHYSICAL ATTACKS ON POWER GRIDS UNDER THE AC MODEL

Daniel Bienstock, Mauro Escobar

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We describe an algorithm for computing undetectable cyber-physical attacks on power grids under the AC power flow model. The adversary's actions affect a small zone of the network; within this zone the adversary can modify demands as well as signals. Both actions are calculated so as to hide the underlying truth, which includes severe equipment overloads, while remaining consistent (i.e. not noticed) from the perspective of the control center of the system. We provide an algorithm and run experiments on large grids.

Introduction

Recent attacks on power grids [1] and extensive blackouts have motivated the study of physical and cyber attacks on these systems. In [13, 14] the authors consider, under the linearized power flow model, that an adversary has the ability of disconnect lines from the network and block that information from the control center (CC) of the grid. An algorithm and conditions, under the AC model, in which these failures can be detected is proposed in [16, 15].

Pure data infection attacks are studied in [5, 7], where an adversary injects false information to the sensors in the network, so that wrong scheduling decisions are made. Also see [12, 10, 8, 9, 11].

In this paper, we consider an adversary that modifies the demands and data over a zone of the grid, so as to hide an overload that results from the demand changes.

Notation. Throughout this document we will use the following terminology: j denotes the imaginary unit $\sqrt{-1}$; for $v \in \mathbb{C}$, v^* denotes its complex conjugate; for a node k, $\delta(k)$ is the set of edges incident to k; for a set $\mathcal{A}$ of nodes of a graph $G = (\mathcal{N}, \mathcal{E})$, let $\mathcal{A}^C \doteq \mathcal{N} \backslash \mathcal{A}$ denote its complement, and let $N(\mathcal{A}) \doteq \{v \in \mathcal{A}^C : \exists u \in \mathcal{A}, uv \in \mathcal{E}\}$ be the neighborhood of $\mathcal{A}$.

Power Flows

A power grid can be characterized by a set $\mathcal{N}$ of nodes (buses) that generate or demand power and a set of branches or transmission lines between the buses, each of these branches has a complex admittance y_{km}. In the AC power flow model, given the demand and generation at each node, underlying physics describe the status of the network through complex voltages $V_k = |V_k| e^{j\theta_k}$ at each bus k, and the complex power flow from bus k to m is given by $S_{km} = V_k (y_{km}(V_k - V_m))^*$. Thus a feasible AC power flow solution must satisfy:

$$\sum_{km \in \delta(k)} S_{km} = S_k^g - S_k^d \qquad \text{for each bus } k, \tag{1}$$

$$V_k^{min} \le |V_k| \le V_k^{max} \qquad \text{for each bus } k, \tag{2}$$

$$|S_{km}| \le S_{km}^{max} \qquad \text{for each line } km. \tag{3}$$

In these expressions, S_k^g and S_k^d represent the generation and the demand at bus k, respectively, S_{km}^{max} is the capacity of branch km, and V_k^{min} and V_k^{max} are lower and upper bounds of the voltage magnitude. Equation (1) states the power balance at bus k must equal the difference between generation and demand.

Finding solutions for AC power flow problems is strongly NP-hard [4] as a result of the quadratic dependence on the voltage of the power flow.

In normal operation of a grid voltage and current are measured periodically at sensors (RTUs and PMUs). Each sensor is located on some branch km close to one of the buses (k or m), and reports thee voltage at this bus and the complex current $I_{km} = y_{km}(V_k - V_m)$ on this branch. These values are reported to the CC.

Attack Model

Assume that an adversary has control over a set of buses $\mathcal{A} \subset \mathcal{N}$, the *attacked zone*, which does not include any generator buses (assumed harder to control). For every bus k in the attacked zone, the adversary has the ability of:

1. modify the bus demand S_k^d,
2. modify each measurement (voltage and current) reported to the CC by any sensor adjacent to k.

The objective of the adversary is to modify the demands within the attacked zone in order to create a large line

overload. The attacker is also modifying the data originating within the zone as to seamlessly present a normal situation (no overload) as far as the CC is concerned; a dangerous condition [3] [1]. Let V_k^T and V_k^R denote the true and reported voltages at bus k, and let $S_k^{d,T}$ and $S_k^{d,R}$ be the true demand of bus k and the demand computed from the reported voltages at bus k, respectively. Then, $(V_k^T, S_k^{d,T})$ needs to solve (1)-(2) and $(V_k^R, S_k^{d,R})$ solves (1)-(3).

In order to obtain undetectability, any sensor located on a branch that connects $\mathcal{A}$ with $\mathcal{A}^C$ must report consistent current and voltages. Thus, the attacker only needs that

$$V_k^T = V_k^R \qquad \text{for each bus } k \in \mathcal{A}^C \cup N(\mathcal{A}^C),$$
$$S_k^{d,T} = S_k^{d,R} \qquad \text{for each bus } k \in \mathcal{A}^C.$$

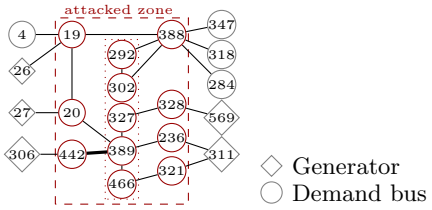


Figure 1: Attack on case2746wp (2746 nodes). The thick line shows the overload.

Experiments

We are able to generate undetectable attacks for large networks on the Matpower case library [18]. The attack on the zone shown in Figure 1 was obtained by solving the adversarial problem with IPOPT [17]. Note the large overload on Table 2.

k	$V_k^T = V_k^R$	k	V_k^T	V_k^R
19	1.090∠-4.96	292	1.110∠-8.22	1.110∠-8.23
20	1.090∠-4.96	302	1.110∠-8.22	1.110∠-8.23
442	1.093∠-11.16	327	1.095∠-10.01	1.095∠-10.03
388	1.111∠-8.23	389	1.104∠-10.02	1.102∠-10.20
328	1.094∠-10.01	466	1.106∠-10.04	1.105∠-10.14
236	1.105∠-10.13	321	1.108∠-10.05	1.108∠-10.05

Table 1: Voltage of subset of attacked buses.

k, m	$\lvert S_{km}^T \rvert$	$\lvert S_{km}^R \rvert$	S_{km}^{max}
389, 442	**143.6**	120.0	120
389, 20	153.8	159.7	160
389, 466	7.2	11.8	120
389, 236	7.0	8.9	120
389, 327	12.4	11.3	120
466, 321	6.8	11.6	120
327, 328	12.5	10.7	120

Table 2: True and reported power flows.

References

[1] Analysis of the cyber attack on the Ukrainian power grid 2016. http://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS _Ukraine_DUC_18Mar2016.pdf, Mar. 2016.

[2] A. R. Bergen and V. Vittal. *Power Systems Analysis*. Pearson, Jan. 2006.

[3] D. Bienstock. *Electrical Transmission System Cascades and Vulnerability: An Oper. Research Viewpoint*. SIAM, 2015.

[4] D. Bienstock and A. Verma. Strong NP-hardness of AC power flows feasibility. *arXiv:1512.07315*, Dec. 2015.

[5] D. Deka, R. Baldick, and S. Vishwanath. Data attacks on power grids: Leveraging detection. In *2015 IEEE Power Energy Soc. Inn. Smart Grid Techn. Conf.*, pages 1–5, Feb. 2015.

[6] J. D. Glover, T. J. Overbye, and M. S. Sarma. *Power System Analysis and Design*. Cengage Learning, 2012.

[7] T. T. Kim and H. V. Poor. Strategic protection against data injection attacks on power grids. *IEEE Trans. Smart Grid*, 2(2):326–333, June 2011.

[8] X. Liu and Z. Li. False Data Attacks Against AC State Estimation With Incomplete Network Information. *IEEE Trans. Smart Grid*, 8(5):2239–2248, 2017.

[9] X. Liu and Z. Li. Local topology attacks in smart grids. *IEEE Trans. Smart Grid*, 8(6):2617–2626, 2017.

[10] X. Liu, Z. Li, and Z. Li. Optimal Protection Strategy Against False Data Injection Attacks in Power Systems. *IEEE Trans. Smart Grid*, 8(4):1802–1810, 2017.

[11] X. Liu, Z. Li, X. Liu, and Z. Li. Masking transmission line outages via false data injection attacks. *IEEE Trans. Inf. Forensics and Secur*, 11(7):1592–1602, 2016.

[12] Y. Liu, P. Ning, and M. Reiter. False data injection attacks against state estimation in electric power grids. *ACM Trans. Inf. Syst. Secur.*, 14:13:1–13:33, 2011.

[13] S. Soltan, M. Yannakakis, and G. Zussman. Power Grid State Estimation Following a Joint Cyber and Physical Attack. *IEEE Trans. Control of Network Systems*, (99), 2016.

[14] S. Soltan, M. Yannakakis, and G. Zussman. REACT to Cyber Attacks on Power Grids. *CoRR*, abs/1709.06934, 2017.

[15] S. Soltan and G. Zussman. EXPOSE the Line Failures following a Cyber-Physical Attack on the Power Grid. *CoRR*, abs/1709.07399, 2017.

[16] S. Soltan and G. Zussman. Power grid state estimation after a cyber-physical attack under the AC power flow model. *2017 IEEE Power and Energy Soc. Gen. Meeting*, pages 1–5, 2017.

[17] A. Wächter and L. T. Biegler. On the Implementation of a Primal-Dual Interior Point Filter Line Search Algorithm for Large-Scale Nonlinear Programming. *Math. Programming*, 106(1):25–57, 2006.

[18] R. D. Zimmerman, C. E. Murillo-Sánchez, and R. J. Thomas. MATPOWER: Steady-State Operations, Planning, and Analysis Tools for Power Systems Research and Education. *IEEE Trans. Power Systems*, 26(1):12–19, Feb. 2011.

[1] The total demand of the network might change when the adversary perform the attack. Our model correctly handles this fact through a model of secondary response [2, 3, 6].

RUMOR SOURCE DETECTION WITH MULTIPLE OBSERVATIONS UNDER ADAPTIVE DIFFUSION PROTOCOLS

Miklos Z. Racz, Jacob Richey

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Recent work, motivated by anonymous messaging platforms, has introduced adaptive diffusion protocols which can obfuscate the source of a rumor: a "snapshot adversary" with access to the subgraph of "infected" nodes can do no better than randomly guessing the entity of the source node. What happens if the adversary has access to multiple independent snapshots? We study this question when the underlying graph is the infinite d-regular tree and show, among other things, that already with three observations there is a simple algorithm that finds the rumor source with constant probability, regardless of the adaptive diffusion protocol.

Introduction and related work

Detecting the source of information diffusion on a network is an important problem in network science, with applications such as finding the source of a virus epidemic or finding the source of a rumor on Twitter. A prototypical graph on which source detection is studied is the infinite d-regular tree $\mathbb{T}_d$ (with $d \geq 3$). For the purposes of this abstract we will focus on this underlying graph, and only briefly mention work on other graphs.

Rumor source detection. Perhaps the simplest and most natural model of information diffusion on a network is the susceptible-infected (SI) model, where the rumor is spread along each edge of the network at a constant rate, and once a node is infected it remains infected forever. Shah and Zaman studied detecting the source in this model [3, 4]. Formally, at time $t = 0$ a vertex $v^* \in \mathbb{T}_d$ is "infected" and the information propagates on the network according to the SI model; one then observes the subset V_t of infected vertices at time t, which consists of $N_t := |V_t|$ vertices. We assume that the underlying graph (in this case $\mathbb{T}_d$) is known and hence the subgraph G_t induced by the vertices in V_t is also known. The goal is to find the rumor source v^*.

Shah and Zaman showed that the maximum likelihood estimator (MLE) $\widehat{v}_{\mathrm{ML}} := \arg\max_{v \in G_t} \mathbb{P}(G_t \mid v^* = v)$ has particularly nice properties in this setting [3, 4]. In particular, they showed that it is computable in linear time and that it detects the source with constant probability. More precisely, they show (in [5]) that there exists a universal constant $\alpha_d > 0$ such that $\lim_{t\to\infty} \mathbb{P}(\widehat{v}_{\mathrm{ML}} = v^*) = \alpha_d$ (when $d \geq 3$). Many results extend to more general settings such as random trees [5].

Wang et al. [6] studied rumor source detection in the same setting but now with multiple independent observations; that is, observing the infected nodes $V_t^{(1)}, \ldots, V_t^{(k)}$ of k independent diffusions started from the same source v^*. They show that the detection probability increases with k and that it goes to 1 exponentially as $k \to \infty$.

Rumor source obfuscation. The results above show that if information propagates according to the SI model, then the source can be found efficiently and with good probability. In certain applications, such as anonymous messaging apps, this is undesirable. Motivated by these applications, Fanti et al. [2] asked whether it is possible to devise messaging protocols that can *obfuscate* the rumor source, while at the same time still spreading information widely and quickly.

They devised a family of messaging protocols, termed *adaptive diffusions*, for this purpose. Their main result shows that a specific messaging protocol within this family achieves *perfect obfuscation*: under this spreading model a "snapshot adversary" can do no better than randomly guessing:

$$\mathbb{P}(\widehat{v}_{\mathrm{ML}} = v^* \mid N_t = n) = \frac{1 + o(1)}{n}.$$

Many results extend to more general settings such as irregular trees [1].

Results

We examine the robustness of spreading algorithms—in particular, adaptive diffusion protocols—in the context of multiple independent observations. We show that when an adversary has access to two observations then a weak form of obfuscation is still possible. However, when it has

access to three or more independent snapshots then source detection with constant probability is always possible, regardless of the adaptive diffusion protocol. We also discuss connections to *local spreading*.

Two independent observations. Only a weak form of source obfuscation is possible when the adversary has two independent observations. This is characterized up to a constant factor in the following theorem.

Theorem 1. *Suppose that information is spread according to an adaptive diffusion protocol and that an adversary has two independent observations of the infected nodes at time t started from a fixed source node v^*.*

(a) There exists an efficient estimator $\widehat{v}$ and a constant $c > 0$, both independent of the spreading protocol, such that

$$\mathbb{P}(\widehat{v} = v^*) \geq \frac{c}{\log(N_t)}. \tag{1}$$

(b) There exists an adaptive diffusion protocol such that

$$\mathbb{P}(\widehat{v}_{\mathrm{ML}} = v^*) \leq \frac{C}{\log(N_t)}. \tag{2}$$

Three or more observations. If the adversary has three or more observations, then the source can always be detected with constant probability. Moreover, this probability converges to 1 exponentially as $k \to \infty$.

Theorem 2. *Suppose that information is spread according to an adaptive diffusion protocol and that an adversary has $k \geq 3$ independent observations of the infected nodes at time t started from a fixed source node v^*.*

There exists an efficient estimator $\widehat{v} = \widehat{v}(k)$, independent of the spreading protocol, such that

$$\mathbb{P}(\widehat{v} = v^*) \geq 1 - e^{-Ck}, \tag{3}$$

where $C > 0$ is a universal constant.

Local spreading. It is often desirable to not only spread information widely and quickly, but also to spread it *locally* around the source. However, this is at odds with obfuscating the source. To understand this tradeoff, we introduce a formal notion of *local spreading*: let R_t denote the radius of the largest ball of infected nodes centered at the rumor source at time t; we define the local spreading of a protocol at time t to be $\mathbb{E}[R_t]$.

The adaptive diffusion protocol that achieves perfect obfuscation does not have local spreading; in this case $\mathbb{E}[R_t] = O(1)$ as $t \to \infty$. Ideally we would like local spreading to grow linearly in t; to achieve this one has to relax the obfuscation requirements. The following theorem characterizes the optimal tradeoff (under a single snapshot) and shows that the optimum is obtained by an adaptive diffusion protocol.

Theorem 3. *Consider a snapshot adversary.*

(a) Suppose that a spreading protocol achieves "polynomial obfuscation", that is, the following holds:

$$\mathbb{P}(\hat{v}_{\mathrm{ML}} = v^*) \leq \frac{1}{N_t^{\gamma}}(1 + o(1)) \tag{4}$$

for some $\gamma \in (0, 1)$ and all $t > 0$. Then

$$\mathbb{E}[R_t] \leq (1 - \gamma + o(1))t/2. \tag{5}$$

(b) For any $\gamma \in (0, 1)$ there exists an adaptive diffusion algorithm satisfying both (4) and

$$\mathbb{E}[R_t] = (1 - \gamma - o(1))t/2.$$

We also discuss how local spreading is related to obfuscation under multiple observations.

Discussion and future work

The main message of this work is that while adaptive diffusion protocols can hide the source from a snapshot adversary, they are ineffective when the adversary has access to multiple independent snapshots. The main question raised by our work is whether there exist other diffusion protocols that can obfuscate the source against such an adversary.

References

[1] G. Fanti, P. Kairouz, S. Oh, K. Ramchandran, and P. Viswanath. Hiding the rumor source. *IEEE Transactions on Information Theory*, 63(10):6679–6713, 2017.

[2] G. Fanti, P. Kairouz, S. Oh, and P. Viswanath. Spy vs. Spy: Rumor Source Obfuscation. In *ACM SIGMETRICS Performance Evaluation Review*, volume 43, pages 271–284. ACM, 2015.

[3] D. Shah and T. Zaman. Detecting Sources of Computer Viruses in Networks: Theory and Experiment. In *ACM SIGMETRICS*, volume 38, pages 203–214. ACM, 2010.

[4] D. Shah and T. Zaman. Rumors in a Network: Who's the Culprit? *IEEE Transactions on Information Theory*, 57(8):5163–5181, 2011.

[5] D. Shah and T. Zaman. Finding rumor sources on random trees. *Operations Research*, 64(3):736–755, 2016.

[6] Z. Wang, W. Dong, W. Zhang, and C. W. Tan. Rumor Source Detection with Multiple Observations: Fundamental Limits and Algorithms. In *SIGMETRICS*, volume 42, pages 1–13. ACM, 2014.

Poster presentations
(in alphabetical order)

A PRINCIPAL STRATIFICATION APPROACH TO UNCOMPLICATE CAUSAL INFERENCE COMPLICATIONS ON SOCIAL NETWORKS

Kristen M. Altenburger

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Randomized experiments in practice are often broken, suffering from complications such as missing outcome data or treatment non-compliance issues. In this ongoing work, we extend the principal stratification framework [6] to address treatment non-compliance [1] specifically in social network settings. We consider a randomized experiment on a fully observed social network where edges represent possible pathways of treatment interference or other interference like knowledge-sharing. We assume a random subset of nodes are first encouraged to share information (as in [4]) at time t, then they either share or don't share, and the real estimand of interest is the spillover effect that has among the node's friends or followers at a later time $(t+1)$[1]. This work's main conceptual contribution is that in network settings, the latent principal strata of nodes that are initially treated defines the subsequent principal stratification of their friends, extending work in [5].

The Principal Stratification Approach

Principal stratification [6] is defined as a partition of units based on a post-treatment variable defined by $S_{i,t}(Z_i)$ where groups of units are defined by the joint value $(S_{i,t}(1), S_{i,t}(0))$, where Z_i represents treatment assignment for unit i. The aim with the traditional principal stratification framework is to define post-treatment strata that permits an analysis on a comparable set of units, where stratum membership is assumed to be unaffected by treatment assignment. By stratifying units based on a post-treatment variable such as treatment uptake, then principal effects can be defined on a common set of units, thus avoiding introducing any post-treatment bias.

Extending Principal Stratification to Social Networks

Consider a randomized experiment run on a social network, where the spillover effect is of interest [3, 4]. For concreteness, suppose that users are encouraged to tweet a link to all their followers [4]; we motivate the need to address complications such as compliance where not every unit follows their treatment assignment. The additional information of the network structure allows one to better account for the subsequent stratification of the follower nodes to estimate spillover effects on a common set of units. We focus on the familiar problem of non-compliance in randomized experiments, though this framework generalizes to other principal stratification set-ups. We outline below the possible stratification among the two sets of nodes we define as: the *broadcasters* are those nodes randomly encouraged to share information with their followers, and the *followers* are those nodes that follow at least one broadcaster node. We use the word "follow" or "friend" interchangeably to refer to a directed relationship between follower i and broadcaster j.

Stratification of Broadcasters: Suppose a seed set of users is randomly encouraged to share information to all their friends. This assignment-to-treatment variable will be encompassed by the treatment vector Z where $Z_i = 1$ if user i is encouraged, and 0 otherwise. We conceptualize the network of interest as a bi-partite network where one node set consists of the broadcasters, and the remaining set consists of nodes who follow the broadcasters, excluding any broadcasters that follow each other. We define $S_{i,t}$ as an indicator for whether broadcaster node i shares information. Then, per the usual set-up in encouragement designs, the principal strata defined among this seed set of broadcasters are:

- BB: $S_{i,t}(1)$=$S_{i,t}(0)$=1, users who would broadcast regardless of treatment assignment [always-takers],
- BN: $S_{i,t}(1)$=1, $S_{i,t}(0)$=0, users who would broadcast if treated and would not otherwise [compliers]
- NB: $S_{i,t}(1)$=0, $S_{i,t}(0)$=1, users who would not broadcast if treated and would otherwise [defiers]
- NN: $S_{i,t}(1)$= $S_{i,t}(0)$=0, users who would never broadcast regardless of treatment status [never-takers]

Stratification of Followers: The set of users not in the broadcaster set and who follow at least one user in the broadcaster set are defined as a follower. We define the

[1] We assume that treatment spillover effects can only happen among one's immediate friends in the network (i.e. 1st-order spillover effects [2]) and not between one's friends-of-friends.

treatment assignment as Z_i^*=1 if user i follows either at least one always-taker/complier or only never-takers that were encouraged to share information and 0 otherwise[2]. We define $S_{i,t+1}^*$ as an indicator for whether a follower node i receives broadcasted information. The non-existence of defiers can be enforced by design or is common to assume does not exist in most settings, and we adopt the same assumption on the broadcaster set here as well. The stratification of follower nodes receiving, R, treatment is:

- RR: $S_{i,t+1}^*(1)$=$S_{i,t+1}^*(0)$=1, users who follow at least one always-taker
- RN: $S_{i,t+1}^*(1)$=1,$S_{i,t+1}^*(0)$=0, users who only follow compliers or compliers and never-takers
- NN: $S_{i,t+1}^*(1)$=$S_{i,t+1}^*(0)$=0 users who only follow never-takers

We provide a schematic illustration and reasoning of the follower stratification shown in Figure 1. There are 3 overall following types to consider as marked by the vertical lines where a user i can follow nodes from 1-3 distinct principal strata. The top row of nodes indicates the principal stratification of the broadcasters at time t. Then the bottom row of nodes indicates the resulting stratification of the follower nodes at time $t+1$. We propose the following explanations for describing how strata membership is propagated to the follower nodes. If user i only follows users that belong to one particular strata, then the follower nodes automatically adopt that same strata membership as well. For example, if user i only follows always-takers, then user i can only be considered an always-taker because they will be exposed to treatment information regardless of the broadcaster's treatment assignment.

If user i follows users that belong to multiple distinct strata, then we justify the following stratification of followers. If user i follows an always-taker and a never-taker, then user i is considered an always-taker. Regardless of the other broadcaster's behavior, user i will always be exposed to the treatment as long as they follow at least one always-taker. Therefore, more generally, whenever user i follows an always-taker, they automatically become an always-taker as well. If user i follows a never-taker and a complier, then we consider them a complier since the complier broadcaster node will determine treatment

[2]We note there are several ways to define Z^*. Here we define Z^* in terms of the intent-to-encourage effect; the assumption is that treatments are equivalent regardless of which follower type it's received from.

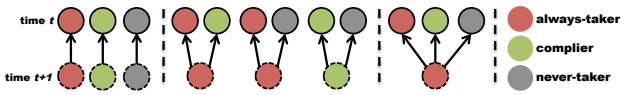


Figure 1: Schematic illustration of the principal strata structure imposed on follwers at time $t+1$.

exposure. Finally, when user i follows other users from all the strata, we again consider i an always-taker since again following at least one always-taker implies i is an always-taker. Then the causal estimand of interest can be defined as τ_g=$\mathbb{E}[Y_{i,t+1}(1)$ - $Y_{i,t+1}(0)$ $|g]$ where $g = RR$, RN, or NN and $Y_{i,t+1}$ denotes the outcome.

In most settings, interest is on understanding the compliers, where we now account for always-taker followers who will always have access to treatment even if they do follow both always-takers and compliers. A complier is now defined a user who only follows compliers, or follows a combination of compliers and never-takers only.

Conclusion

The benefit of formulating spillover complications in the language of principal stratification is that it provides a more general way to address other post-treatment complications. For example, principal strata could be defined by post-treatment knowledge under both treatment educational assignments, assuming that having knowledge is a necessary condition to be able to share treatment with others [3]. This proposed extension of the traditional principal stratification framework begins to address complications for randomized experiments on social network settings so that causal effects can be defined on a more comparable set of units.

References

[1] J. D. Angrist, G. W. Imbens, and D. B. Rubin. Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455, 1996.

[2] S. Athey, D. Eckles, and G. W. Imbens. Exact p-values for network interference. *Journal of the American Statistical Association*, pages 1–11, 2017.

[3] J. Cai, A. De Janvry, and E. Sadoulet. Social networks and the decision to insure. *American Economic Journal: Applied Economics*, 7(2):81–108, 2015.

[4] A. Coppock, A. Guess, and J. Ternovski. When treatments are tweets: A network mobilization experiment over Twitter. *Political Behavior*, 38(1):105–128, 2016.

[5] D. Eckles, R. F. Kizilcec, and E. Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016.

[6] C. E. Frangakis and D. B. Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

STREET NETWORK MODELS AND PLANARITY IN URBAN FORM STUDIES

Geoff Boeing

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Urban researchers often use planar graphs to model street networks as a tractable simplification of real-world complexity, which can be useful but can also impact the results of real-world street network analysis. This study measures the nonplanarity of central street networks in 50 cities worldwide. It develops two new indicators of spatial planarity and empirically quantifies how planar models can inconsistently but drastically misrepresent intersection density, street lengths, routing, and connectivity.

Planar Street Network Models

A planar graph can be drawn on a two-dimensional plane without any of its edges crossing each other, except where they intersect at nodes. If it cannot be drawn to meet this criterion, it is nonplanar. Street networks are embedded in space, which provides them with geometry—such as geographical coordinates, lengths, areas, shapes, and angles—along with their topology [1]. Thus we must distinguish between a graph's mathematical/topological planarity, which we refer to as *formal planarity*, and the planarity of its real-world spatial embedding, which we refer to as *spatial planarity*. For example, a street network might be spatially nonplanar due to its embedding in space but it could still be formally planar. That is, if we redraw the graph (i.e., altering its geometry without altering its topology), there may exist some other embedding that prevents edges crossing anywhere but at nodes.

Imposing a planar model on a street network forces nodes at any line crossings. At a citywide or regional spatial scale, nonplanar edge crossings may be relatively uncommon: we might call such a network *approximately* planar [4]. If street networks can be sufficiently well-modeled by planar graphs, there are certain methodological benefits to doing so as they offer computational simplicity and algorithmic tractability. Although planar models can be computationally useful, real-world street networks often include at least one overpass or underpass that results in the failure of formal proofs of planarity. But debating the semantics of formal planarity may be missing the point—more interesting is the "extent" to which a network is nonplanar [5].

The graph theory literature offers some measures of how far-off a nonplanar graph G is from being planar, including its *crossing number* and *skewness*. However, these fail to adjust for the size, density, or real-world embedding of a spatial network. To date, no widely-accepted measure of the extent of nonplanarity has emerged.

Methods

This study empirically analyzes one-square-mile drivable and walkable street networks at the centers of 50 cities worldwide to quantify how planar models impact street network analysis results (focusing on the common urban form measures of intersection density and street segment length) and the extent to which bias varies across places and types of urbanization.

To acquire these street networks, we use OSMnx to download the data for each city and network type from OpenStreetMap. OpenStreetMap is a collaborative online mapping platform commonly used by researchers because of its high-quality worldwide coverage. OSMnx is a tool that allows us to download a street network from OpenStreetMap for any study site in the world, and process it into a length-weighted nonplanar directed graph [2, 3]. It differentiates between walkable and drivable routes based on individual elements' metadata that describe how the route may be used.

We test each network for formal planarity, then calculate two measures of the extent of (spatial) planarity. The first is the Spatial Planarity Ratio, ϕ. It represents the ratio of nonplanar intersections i_n to planar intersections i_p:

$$\phi = \frac{i_n}{i_p} \tag{1}$$

The second is the Edge Length Ratio, λ. It represents the ratio of the planar graph's mean edge length l_p to the nonplanar graph's mean edge length l_n:

$$\lambda = \frac{l_p}{l_n} \tag{2}$$

Street network models almost always impose an artificial boundary on the network, and moving the study site even slightly might affect these results. We explore how these indicators vary throughout a single city by analyzing 100 random samples of the drivable street network of Oakland, California.

Findings

Among these drivable street networks, only 20% are formally planar. On average, they are 88% spatially planar by the ϕ measure and 89% by the λ measure. The individual ϕ values indicate that spatial planarity ranges from a high of 100% in six of these cities to a low of 54% in central Moscow. The λ values indicate that spatial planarity ranges from a high of 100% in six of these cities to a low of 60% in Moscow. On average across these networks, planar models overcount intersections by 16% and underestimate street segment lengths by 11%.

Among walkable street networks, only 10% are formally planar. On average, they are 92% spatially planar by the ϕ measure and 91% by the λ measure. The individual ϕ values indicate that spatial planarity ranges from a high of 100% in two cities to a low of 66% in central Shanghai. The λ values indicate that spatial planarity ranges from a high of 100% in two cities to a low of 64% in Shanghai. On average across these networks, planar models overcount intersections by 10% and underestimate street segment lengths by 9%. Fewer walkable than drivable networks are formally planar, but on average these walking networks are slightly more spatially planar than the driving networks.

Not all formally planar street networks are spatially planar. For example, central Toronto's drivable network is formally planar but only 92% (ϕ) and 95% (λ) spatially planar. In total, four central drivable networks (Toronto, Jakarta, Florence, and Copenhagen) and three central walkable networks (Dallas, Delhi, and Bologna) are formally planar but spatially nonplanar. About a third of the city centers studied demonstrate ϕ spatial planarity of 96% or higher, suggesting that they are approximately planar. However, another third of the city centers are less than 87% planar. Central Dallas, Los Angeles, and Moscow have ϕ values below 59%, suggesting planar graphs poorly model these city centers: planarity overstates the intersection counts in these three networks by 71%, 72%, and 85% respectively.

Mogadishu is the only city studied that demonstrates perfect planarity across all three indicators for both network types. Cities like Milan and Barcelona demonstrate perfect planarity in their centers' drivable networks, but not in their walkable networks. Further, the extent of nonplanarity is not consistent across network types: Dallas's walkable ϕ is 65% greater than its drivable ϕ, while Geneva's drivable ϕ is 15% greater than its walkable ϕ. While nearly every European city is in the highest tercile, indicating their networks are more planar, most American cities are in the lowest tercile, indicating their networks are more nonplanar.

Oakland's city-wide street network is formally nonplanar. In terms of spatial planarity, the city has a ϕ of 91.6% and a λ of 92.9%. This suggests that the planar representation of Oakland's drivable street network overstates the number of intersections—and thus, the network's connectivity—by 9.2% city-wide and understates the average edge length by 7.1% city-wide. However, these indicators vary across the city. The samples' mean ϕ and λ scores are reasonably close to the city-wide values, but the samples range from spatial planarity lows of 56.9% (ϕ) and 56.7% (λ) to highs of 100%. 67% of the samples pass the formal planarity test, but 63% of the samples are at least somewhat spatially nonplanar (i.e., $\phi < 1$).

Conclusion

To summarize, planarity varies both across cities as well as across different neighborhoods within individual cities. We find that planar simplifications can impact modeling and analysis in several ways: intersection counts are overestimated due to false nodes where grade-separated edges cross; average edge lengths are underestimated; and connectivity is misrepresented for routing, accessibility analysis, and topological studies. These impact our understanding of the urban form's density, grain, pattern, connectedness, and permeability.

References

[1] M. Barthelemy. Spatial networks. *Physics Reports*, 499(1-3):1–101, 2011.

[2] G. Boeing. OSMnx: New Methods for Acquiring, Constructing, Analyzing, and Visualizing Complex Street Networks. *Computers, Environment and Urban Systems*, 65:126–139, 2017.

[3] G. Boeing. A Multi-Scale Analysis of 27,000 Urban Street Networks. Under review: https://ssrn.com/abstract=2943038, 2018.

[4] R. Louf and M. Barthelemy. A typology of street patterns. *Journal of The Royal Society Interface*, 11(101):1–7, 2014.

[5] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford, 2010.

QUANTIFYING HUMAN PRIORS OVER GRAPHICAL REPRESENTATIONS OF TASKS

Gecia Bravo Hermsdorff

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Our ability to efficiently solve *new* tasks is sharply constrained by our brains' prior beliefs about their structure. Unfortunately, quantifying such priors is a formidable goal, if not only for the reason that "what is a task?" is a relatively open-ended question. Here, we focus on tasks that have a natural mapping to graphs, and develop a method to quantify human priors over these "task graphs", which combines new modeling approaches with Markov chain Monte Carlo with people [1] (a process whereby an agent learns from data generated by another agent, recursively). We also show that our method recovers priors more accurately than the standard approach. Moreover, we propose a novel low-dimensional "smooth"[1] parametrization of probability distributions over (non-isomorphic) graphs with the same vertex set. We show that, in the limited data regime, it allows for more accurate recovery of the prior (*in silico* data), and better generalization (in human data, which we acquired using our new online experiment platform that gamifies our MCMCP algorithm and allows subjects to interactively draw the task graphs[2]). Finally, we extend our framework to the more general case of quantifying priors over exchangeable random structures.

Markov chain Monte Carlo with people (MCMCP)

Fig. 1 illustrates our algorithm for generating experiments. Specifically, our experiments focus on navigation and social interaction tasks. E.g., the subject is told that they are visiting a new city, and whether certain pairs of neighborhoods share a border or not (*step 1* in fig.1). They then are asked to guess if the other pairs of neighborhoods share a border or not (*step 2*) by drawing this city neighborhood map using our graph drawing interface.[3]

The back-and-forth between data seen by the subjects (relations shown, or "partial graphs") and the resulting hypothesis they infer (completed task graphs) can be marginalized over the partial graphs to create a Markov chain (MC) over the space of task graphs. Assuming that subjects are Markovian, and share the same fixed decision rule, this MC is time-homogeneous. If, in addition, we assume that they are Bayesian and respond by sampling from their posterior,[4] this MC has as its stationary distribution the subjects' shared prior over task graphs. Precisely, this "MCMCP Bayesian model" gives a transition matrix $\underline{\underline{\mathbf{T}}}$ over the relevant non-isomorphic task graphs, with entries:

$$t_{ij} = p(g_i|g_j) = \sum\nolimits_k p(g_i|d_k)p(d_k|g_j) \qquad (1)$$

where $p(d_k|g_j)$ is the probability of seeing partial graph d_k by randomly obscuring r relations of the graph g_j, and $p(g_i|d_k)$ is given by Bayes rule using a fixed prior.

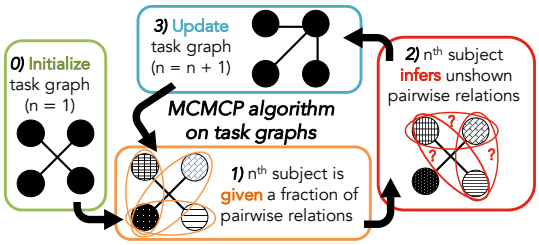


Fig. 1. *Schema of our algorithm for generating experiments.*

Resource constrained MCMCP

In standard MCMC, one uses samples generated by the algorithm to reconstruct the target (stationary) distribution. This is inefficient in the following sense: to obtain i.i.d. samples, only a small fraction of the iterations are used as samples as one must discard the initial samples until (hopefully) the chain has converged, and (to mitigate correlations) can only collect samples every $\mathcal{O}(\tau)$ iterations. While this might not always be a problem (e.g., when samples are efficiently generated via a computer), in MCMCP the primary bottleneck is due to the use of human subjects. Fortunately, here, we can use data more

[1] In the sense that graphs that differ by fewer edges are given similar probabilities.

[2] Links for some of our experiments:
http://psiturk-geciah.princeton.edu:9003 (navigation in cities),
http://psiturk-geciah.princeton.edu:9002 (friendship in classes).

[3] To incentivize subjects to give their true prior, they are told that there is an underlying truth (e.g. an actual city), and that they win extra money by correctly guessing the relations obscured.

[4] Specifically, we assume that the subjects share the same prior, are aware that the "partial graphs" are generated by randomly erasing a fraction of the relations, and that the MC is ergodic.

efficiently by leveraging the additional structure provided by the Bayesian assumption. Specifically, we propose to recover subjects' prior by fitting their choices to our full MCMCP Bayesian model (as opposed to using the graph frequency as a proxy of the prior as it is done in classic MCMC). The unknowns are the probabilities that the prior gives to each of the non-isomorphic graphs. As illustrated in fig. 2, this fitting method recovers the prior more precisely than the standard MCMC sampling method (especially in the case of constrained chain length). Our approach has additional advantages: it does not have the problem of "guessing" the mixing time; and it allows for experiments to be also run in parallel.

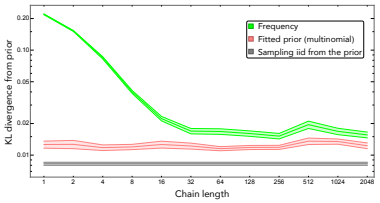


Fig. 2. *Priors can be more precisely recovered by leveraging the MCMCP assumptions.* We simulate data from our MCMCP Bayesian model on 5 nodes (34 non-isomorphic graphs), with a prior chosen to give an exponential mixing time of $\tau_m \sim 13$ iterations. Each simulation has 2048 samples, split into different chain lengths. We fit a multinomial prior to these data. We then calculate the KL divergence from the true prior for the fitted prior and for the sampled frequency of graphs. Error bars denote ± 1 standard error.

Low-dimensional representation of priors

The number of non-isomorphic graphs $G(n)$ on n nodes grows superexponentially; given limited data, even for moderate n we cannot sufficiently sample each graph. For these cases, to obtain informative priors, we need to extend the probabilities to graphs that were not sampled. Our approach is to find a natural low-dimensional parameterization of the prior. Specifically, we propose to use the following form: $\underline{\mathbf{p}} \propto \mathrm{ER}(1/2) * \exp \sum_{b=2}^{G(n)} c_b \underline{\mathbf{v}}_b$, where c_b are the coefficients to be fit, $\underline{\mathbf{v}}_b$ is the b^{th} left eigenvector (ordered by decreasing eigenvalues) of the transition matrix $\underline{\underline{\mathbf{T}}}$ from eq. 1, with the data generated by obscuring only one relation of the underlining graph, and using an ER(1/2) (Erdős-Rényi with $p = 1/2$) distribution as the prior. This choice has several interesting properties,[5] e.g., when $c_{i>2} = 0$, there is a unique correspondence between $c_2 \in (-\infty, \infty)$ and an $ER(p)$ prior with $p \in (0, 1)$. "Smoother" priors are obtained by including only the longer-decaying modes.

As illustrated by the simulations in fig. 3A, when data are limited, using a low-dimensional parametrization recovers the prior more accurately. It also results in better generalization in the human data (fig. 3B). In both cases, as the amount of data increases, using the full basis set does best. Moreover, the fact that the human data are better fit by a full basis set suggests that there is non-trivial graphical structure in subjects' priors.

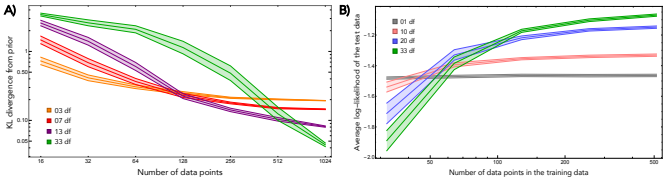


Fig. 3. *Benefits of our low-dimensional smooth parametrization of the prior in the limited data regime.* **A)** *Improved accuracy.* We simulate data using the same specifications as in fig. 2 (with chain length 16), and fit the prior with different numbers of bases (degrees of freedom, df). **B)** *Better generalization.* We use 1210 data points from a single cover story on 5 nodes, randomly split them into test (698 samples) and training data, and fit the prior with different numbers of df.

MCMCP over exchangeable random structures

We extend our results to the more general case of quantifying priors over exchangeable random structures, where the partial data are generated by randomly obscuring a given fraction of the sequence. The relevant parameters are: $\mathcal{A}$, the alphabet; ℓ, the string length; m, the number of relations obscured; and $\mathcal{G}$, the group under which the sequence is exchangeable.[6] An element in $\mathcal{G}$ induces a permutation of the indices $\{1, \ldots, \ell\}$, and thus a permutation of the elements in $\mathcal{A}^\ell$. This action of $\mathcal{G}$ induces an equivalence relation on $\mathcal{A}^\ell$ ($x \sim y$ if $\exists g \in \mathcal{G}$ s.t. $x = g.y$), which partitions it into equivalence classes.[7] The condition of exchangeability under $\mathcal{G}$ means that probabilities are assigned to these partitions, with elements in the same partition having equal probability.

References

[1] T. Griffiths & M. Kalish. Language evolution by iterated learning with Bayesian agents. *Cognitive Science*, 31:441–80 (2007).

[5] Formal details about this parametrization and its extension to exchangeable random structures to appear in a paper under preparation by Bravo Hermsdorff, G. & Gunderson, L.M.

[6] E.g., for unordered binary strings, $\mathcal{A} = \{0, 1\}$ and $\mathcal{G}$ is the *full* permutation group $\mathcal{S}_\ell$ acting on the entries of strings of length ℓ. For simple graphs, the binary string is length $\ell = \binom{n}{2}$ and $\mathcal{G}$ is the permutation group $\mathcal{S}_n$, where n is the number of *nodes*.

[7] For unordered binary strings, they are partitioned into $\ell + 1$ sets, one for each possible sum $(0, ..., \ell)$; for simple graphs, the partitions correspond to the non-isomorphic graphs on n nodes.

A UNITARY OPTIMISATION PROCESS FOR THE FORMATION OF HETEROGENEOUS STRUCTURES IN BIPARTITE MUTUALISTIC NETWORKS

Weiran Cai, Jordan Snyder, Raissa D'Souza and Alan Hastings, UC Davis

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We discuss a unitary dynamical process that generates the dual structural properties observed in ecological and socio-economic mutualistic networks. A mechanism of dyadic preferential attachment through population and niche structure is recognised.

Abstract

Mutualistic networks are critical components of ecological systems [1]. Empirical studies found that species in them have a salient inclination of being simultaneously nested and modular in their reciprocal relations [2]. Whether and how the two link patterns can emerge from a unitary mechanism however remains unclear. Most proposed models concern only the aspect of nestedness while few are on the cause of modularity. We propose an optimisation model that reproduces both structural features. We utilise population dynamics with a niche structure. Niche overlaps are interpreted as the intensities of involved mutualistic and competitive interactions. A mechanism of dyadic preferential attachment through a positive feedback of population heterogeneity and link centrality is recognised in the network formation. Off the constraints of static models, we further examined the network's distinctive dynamical properties at various time scales, from local stability and reversibility, to niche evolution and extinction. Most notably, a profound history-dependency of the structural properties is revealed on environmental alterations at a large time scale. The principle, when extended to payoffs of other specialties, could also underlie the formation of socio-economic networks that exhibit bipartite cooperation [3].

Meeting hashtag: #SIAMNS18

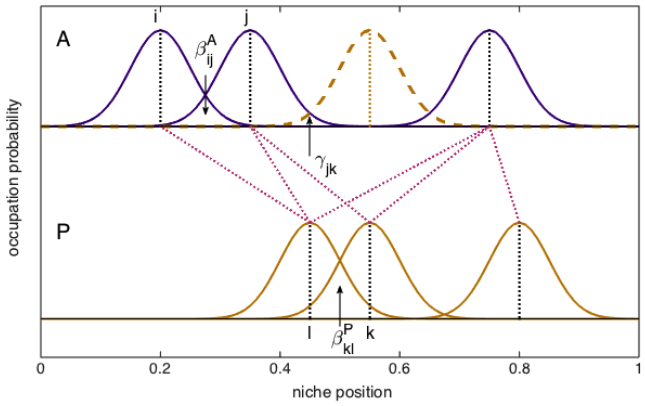


Figure 1: Optimisation model built on niche structure.

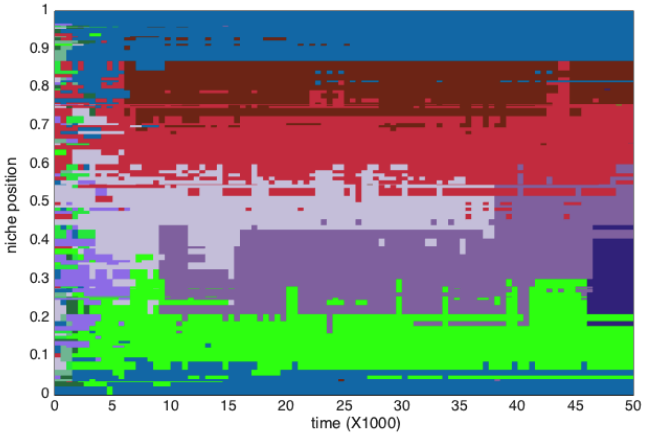


Figure 2: Time course of modularization.

References

[1] Bastolla, U., et al., *The architecture of mutualistic networks minimizes competition and increases biodiversity*. Nature 458(23), 07950 (2009).

[2] Olesen, J. M., Bascompte, J., Dupont, Y. L., and Jordano, P. *The modularity of pollination networks*. PNAS 104(50), 19891 (2007).

[3] Saavedra, S., Reed-Tsochas, F., and Uzzi, B. *A simple model of bipartite cooperation for ecological and organizational networks*. Nature 457(22), 463 (2009).

BACKBONE STRUCTURE OF HIERARCHICAL NETWORK PARTITIONING

Zizhen Chen, David W. Matula

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Extended Abstract

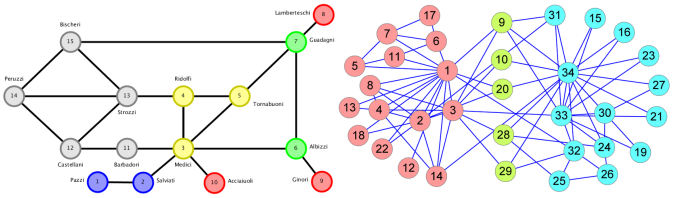
We review the foundations of a hierarchical network partitioning method based on the fundamental concept of density (i.e., sparse cuts separating dense components) and implemented via a duality to peer-to-peer network flow between all node pairs. The methodology is conceptualized by considering traffic flow in a road network with traffic demands between all cities. In periods of high flow "congestion pricing" might be imposed to lessen flow through a bottleneck set of roads, such as those into a city's downtown core. For regions of somewhat less congestion successively smaller prices might be imposed across boundary accesses. To determine this hierarchy of boundaries we employ the Maximum Concurrent Flow Problem (MCFP) which can be formulated as a Linear Program (LP) with maximin objective to maximize the minimum flow throughput guaranteed between all node pairs subject to the paths with flow sharing capacity of the edges [1, 7, 9, 10, 11, 12, 13]. Throughput is the ratio of the flow delivered between a node pair in comparison to the node pair's corresponding demand. Employing LP duality, the optimal throughput is shown to determine a critically saturated separating set of edges partitioning the network into component parts, where all edges of the components have strictly positive residual capacity.

The hierarchical MCFP is then formulated to further maximize throughput in the residual capacity components determining a second throughput level and set of critical edges. Iterating further employing a series of LP's, a series of throughput levels is determined until all edges are critical, yielding a stratification portrayed in hierarchical partitioning theory as a dendrogram.

Each LP of the HMCFP sequence of LP's determines in polynomially bounded time either a sparsest cut or a sparse grid, the latter a multipart partition proven to have at least five parts. The result is consistent with the sparsest cut problem by itself being NP-hard [8, 12, 6].

Most real world networks from many fields are sparse. Numerous examples having a power law distribution of degrees have been cited [3, 2], but the generality of this observation is problematic [5]. Importantly, we found that networks having ground truth community structure typically yield successive levels of marginally sparsest cuts before encountering a gridlock fragmentation. We utilized well-known small social networks (Figure 1) to illustrate the process and exhibit the component communities by characterizing the "backbone" concept of the network.

In Figure 1a, the 15-node Florentine Families network [4] is represented as a graph. Taking edge capacities and node pair demands both as unity provides a density based hierarchy by the HMCFP. The backbone can be defined as a "graph minor", with the components at any level formed by contraction of the node sets.



(a) Florentine Family Network (b) Karate Clube Network

Figure 1: Two Small Networks Illustrating the Procedure

An edge of the contracted graph after k cuts can be labeled by the set of j cuts that include that edge $1 \leq j \leq k$. A path of two or more edges between the end nodes of an edge that is cut by the same set of j cuts is termed a "back channel" between the nodes. The edges with no back channels between their end nodes are "backbone edges" and characterize the subgraph of the contracted graph termed a "backbone". Backbone edges provide the excess capacity to absorb the additional flow between end node pairs to maximize the concurrent flow at that level.

The minimal set of cuts crossed by a path between two components of the backbone provides a distance between the components. The concept of distance between components identifies weak and strong relations between the components. For community detection applications, this adds to the understanding of relations between communities augmenting the original data on relations between

individual node pairs. Figure 2 shows the Florentine Family backbones from cut level 1 to level 5.

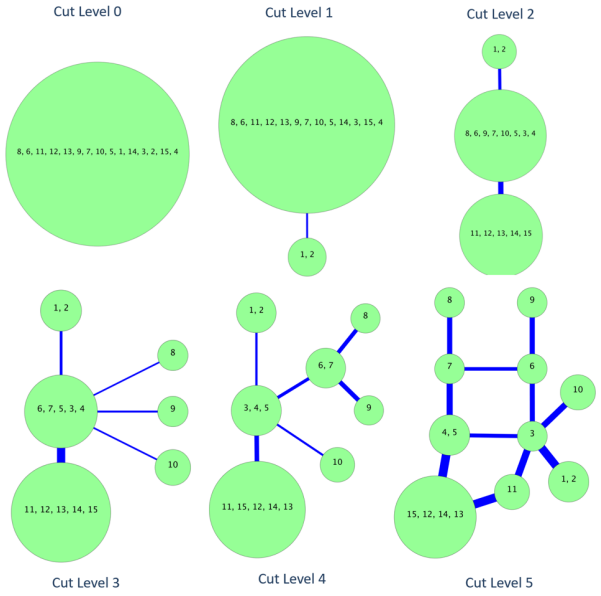


Figure 2: Florentine Backbones Level 1 to Level 5

The backbone shows the structure of the network by preserving the minimal set of critical edges that guide the flow routing. It is typically a sparse planar graph for real world networks. Figure 1b shows a denser social network: Zachary's Karate Club [14] containing 34 nodes with 78 edges. Figure 3 is the Karate Club backbone at a final cut level before gridlock occurs. The backbone provides a simplified visualization of the denser components and their interrelations. Besides the hierarchical component relational structure, the backbone may also provide a seriation of the corresponding adjacency matrix. Figure 4a shows a seriated adjacency matrix compared to Figure 4b which is a randomized adjacency matrix of the Karate Club network.

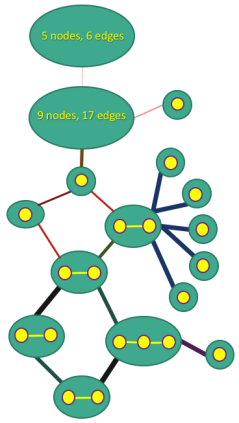
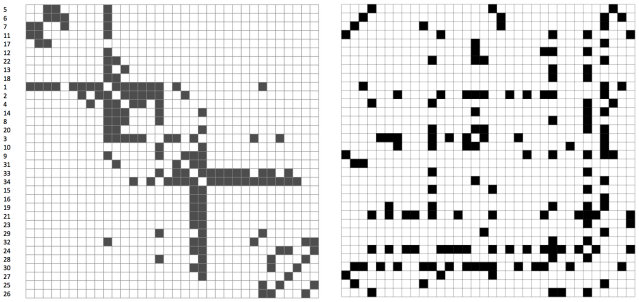


Figure 3: Karate Club Network Final Backbone



(a) Seriated Matrix (b) Random Matrix

Figure 4: Adjacency Matrices of Karate Club Network

References

[1] M. Allalouf and Y. Shavitt. Maximum flow routing with weighted max-min fairness. In *Quality of Service in the Emerging Networking Panorama*, pages 278–287. Springer, 2004.

[2] A. Barabási and M. PÃ3sfai. *Network Science*. Cambridge University Press, 2016.

[3] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[4] R. L. Breiger and P. E. Pattison. Cumulated social roles: The duality of persons and their algebras. *Social networks*, 8(3):215–256, 1986.

[5] A. D. Broido and A. Clauset. Scale-free networks are rare. *arXiv preprint arXiv:1801.03400*, 2018.

[6] S.-W. Chiou. A combinatorial approximation algorithm for concurrent flow problem and its application. *Computers & operations research*, 32(4):1007–1035, 2005.

[7] Y. Dong, E. V. Olinick, T. Jason Kratz, and D. W. Matula. A compact linear programming formulation of the maximum concurrent flow problem. *Networks*, 65(1):68–87, 2015.

[8] E. Gourdin. A mixed integer model for the sparsest cut problem. *Electronic Notes in Discrete Mathematics*, 36:111–118, 2010.

[9] T. Leighton and S. Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM (JACM)*, 46(6):787–832, 1999.

[10] D. W. Matula. Cluster validity by concurrent chaining. In *Numerical Taxonomy*, pages 156–166. Springer, 1983.

[11] D. W. Matula. Divisive vs. agglomerative vverage linkage hierarchical clustering. In W. Gaul and M. Schader, editors, *Classification as a Tool of Research*. North-Holland, 6 1985.

[12] D. W. Matula and F. Shahrokhi. Sparsest cuts and bottlenecks in graphs. *Discrete Applied Mathematics*, 27(1-2):113–123, 1990.

[13] F. Shahrokhi and D. W. Matula. The maximum concurrent flow problem. *Journal of the ACM (JACM)*, 37(2):318–334, 1990.

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

PERSONALIZED PAGERANK, LOCAL CLUSTERING, AND THE STOCHASTIC BLOCKMODEL

Fan Chen, Yini Zhang, Karl Rohe

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Introduction

Given a seed node of interest in a large graph, local graph clustering aims to find a set of nodes around the seed that are highly similar to it. By examining only nodes near the seed node, the running time of such an algorithm should be nearly linear in the size of the output, which is much smaller than the entire graph. As such, these techniques are particularly useful for sampling and studying massive graphs.

Spielman and Tang [5] propose ranking the nearby nodes with the landing probability of a random walk starting from the seed node. This algorithm guarantees a slightly worse than the optimal conductance provided by the Cheeger's inequality. In follow-up work, Andersen et al. [1] use the personalized PageRank (PPR) vector to rank the nodes around the seed node, improving the algorithm.

To fully understand the behaviors of these algorithms and to identify the conditions under which they reveal the latent clustering structure, we propose studying the behavior of PPR under a statistical model. A natural way to formulate the local clustering algorithm is under the Stochastic Blockmodel (SBM) [2], treating the block to which the seed node belongs as the desired local cluster. To this end, Kloumann et al. [4] show that using a PPR vector is asymptotically equivalent to the optimal linear discriminant analysis under the SBM, assuming a symmetry condition on the block structure.

This paper uses a different proof technique that allows the characterization of the PPR vector in a wider range of cases. This more general technique allows for different block sizes, more than two blocks, degree heterogeneity within and across blocks, and directed edges. Adopting the Degree-Corrected Stochastic Blockmodel (DC-SBM) [3], we provide a representation of the PPR vector that demonstrates (1) what PPR is doing and (2) why it is biased. In particular, given a seed node in block 1, we show that the cluster returned by the nodes with the largest PPR values contains nodes outside of block 1, except under strong conditions. Then we look into a simple adjustment that is proposed and studied in [1]. We show that it removes the bias from heterogeneous node and block degrees, and establish several optimality results of this method.

Preliminaries

The DC-SBM is an extension of the SBM, allowing node degree heterogeneity [3]. In a DC-SBM with N nodes, each node belongs to one of K blocks. Let $z_v \in \{1, \dots, K\}$ be the underlying block membership of v-th node, for $v = 1, ..., N$. The DC-SBM is parameterized by (1) a $K \times K$ matrix B with $B_{ij} > 0$ for any $i, j = 1, 2, ..., K$, and (2) a series of parameters ($\theta_v > 0$ for every node v) controlling the node degrees. The presence of each edge corresponds to an independent Bernoulli random variable, which depends on the block membership z_v and the node degree parameter θ_v of two end nodes,

$$\mathbb{P}(\text{edge between } u \text{ and } v) = \theta_u \theta_v B_{z_u z_v}.$$

The personalized PageRank vector is a probability distribution on the nodes of the graph that is defined via a random walk on the graph. This construction makes it particularly easy to sample nodes from the PPR distribution. At each step, the random walk either starts over from the seed node with a constant probability (*teleportation constant*) or continues the random walk from the current node. The personalized PageRank vector $p \in [0, 1]^N$ is the stationary distribution of this process, that is, p_v is the probability of the random walk landing on node v at any time. As such, PPR can be used to sample the local cluster of a seed node in a network, in the way that the nearer nodes are more likely to be sampled.

Results

A. Personalized PageRank as a sampling scheme

Our work decomposes the PPR vector $p \in \mathbb{R}^N$ under the population (expectation) DC-SBM as follows,

$$p_v = \theta_v \tilde{p}_{z_v}, \tag{1}$$

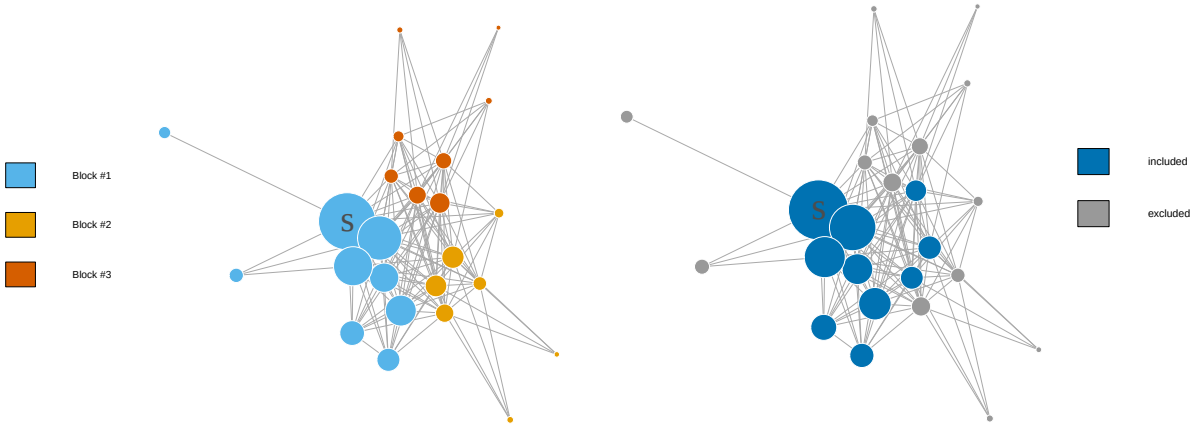


Figure 1: Example of returned local cluster around seed node (marked by "S") using the personalized PageRank vector. The network involves three equally sized blocks with identical block degrees. For both figures, the node size represents the nearness score according to the PPR vector. The colors represent: (left) the true underlying block structure, (right) the nodes included in the returned local cluster (others are colored gray). From the right panel, we observe that using PPR without adjustments yields a biased local cluster.

where $\tilde{p} \in \mathbb{R}^K$ is the "block-wise" personalized PageRank vector that comes from performing PPR on the K node graph with weighed adjacency matrix B.

Based on Equation (1), the paper also extends the previous results of asymptotic equivalence between personalized PageRank and linear discriminant (LD) analysis under some particular DC-SBM [4]. Equation (1) is a population result. Sampling results (not presented due to space constraints) show that the PPR vector concentrates around this population (expectation) under certain conditions.

B. The optimality of personalized PageRank for local clustering under the DC-SBM.

Equation (1) identifies the source of bias of using a personalized PageRank vector for local clustering – the ancillary effects of heterogeneous node degrees θ. The paper shows an additional source of bias due to heterogeneous block degrees (volumes).

Upon this finding, the paper examines a simple adjustment that remedies two biases simultaneously and suggests an appropriate choice of the teleportation constant for which using the PPR vector returns a correct local cluster. In sum, the paper demonstrates that using the adjusted personalized PageRank achieves the precise identification of local cluster, provided optimal sample complexity (up to some constant factor).

Application

Personalized PageRank provides a feasible sample scheme to sample the local cluster in a massive network. By applying the algorithm to the Twitter following graph and sampling the neighborhoods of a few media outlets, we find that using the PPR vector without adjustment returns a local cluster containing primarily popular nodes with high in-degrees on Twitter, whereas adjusted PPR returns noisy results by picking up low in-degree and unrelated handles. We propose a regularization step for the adjustment of PPR and demonstrate its effectiveness in reducing noise.

References

[1] R. Andersen, F. Chung, and K. Lang. Local graph partitioning using pagerank vectors. In *Foundations of Computer Science, 2006. FOCS'06. 47th Annual IEEE Symposium on*, pages 475–486. IEEE, 2006.

[2] P. W. Holland, K. B. Laskey, and S. Leinhardt. Stochastic blockmodels: First steps. *Social networks*, 5(2):109–137, 1983.

[3] B. Karrer and M. E. Newman. Stochastic blockmodels and community structure in networks. *Physical Review E*, 83(1):016107, 2011.

[4] I. M. Kloumann, J. Ugander, and J. Kleinberg. Block models and personalized pagerank. *Proceedings of the National Academy of Sciences*, page 201611275, 2016.

[5] D. A. Spielman and S.-H. Teng. Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems. In *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*, pages 81–90. ACM, 2004.

SIMULATION OF THE NETWORK TOPOLOGY OF GRAIN GROWTH IN METALS

Amanda K. Criner, Austin R. Gerlt, Eric J. Payton, and Jeff P. Simmons

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We detail preliminary results simulating grain growth in metals using random processes on a network. This method accommodates the random nature of grain growth events while considering the evolution of the network topology of the grain structure.

Abstract

Controlling microstructure during manufacture of structural metallic components remains an important concern in many industries. Numerous parts of aircraft and automobiles rely on empirical knowledge of processing parameters to achieve required mechanical properties. The properties result directly from the size, shape, and spatial distributions of grains and phases that compose the microstructure. Concerns over microstructure-related performance often delays the adoption of new alloys and manufacturing processes. Improved multi-scale physics-based simulations of microstructure evolution are a key component for accelerating the insertion of new materials technologies into engineered products [3].

Microstructural evolution metals processing needs to be precisely controlled. There has been significant progress in recent decades in both the simulation and characterization of materials microstructure; however, challenges still remain in translating advances in understanding of nano- and micro-scale materials behaviors to the much coarser scale required for simulation of microstructure variations over an entire manufactured part. Direct simulation methods (such as phase field [7] and Monte Carlo [6]) are capable of excellent agreement with experimental observations; however, they are too computationally intensive to be incorporated into common continuum materials processing simulations. Moreover, the direct simulation techniques are too tedious to be performed on a routine basis. While mean field techniques (in which each grain or particle class in the microstructure is assumed to have similar surroundings) can achieve the desired computational speeds and their input data can be easily determined from conventional metallographic cross-sections, they do not predict the evolution of real microstructures [5, 1]. The scientific community has yet to adequately address how the early stages of grain coarsening affect the spatial arrangement of microstructure, though it may have a significant effect on final grain sizes [1, 8].

We propose a model with the grain structure evolving in time according to an exponential distribution. Two changes of the network of grain boundaries are allowed: a neighbor switch and a grain elimination, as depicted in Figure 1 (a) and (b), respectively. We use total energy change caused by these network transitions, a similar approach to the deterministic models considered in [2, 4, 9], as the exponential rate in our simulations.

We compare the efficiency and flexibility of this approach to other approaches. We discuss the conditions where the obtained distribution demonstrate the self similarity characteristic of normal grain growth, and our future work to accommodate the time-temperature-pressure profiles used during metals processing and predict anomalous, detrimental grain growth through the probability of cascading events in the network.

References

[1] G. Abbruzzese, I. Heckelmann, and K. Lücke. Statistical theory of two-dimensional grain growth–I. The topological foundation. *Acta metallurgica et materialia*, 40(3):519–532, 1992.

[2] A. Cocks and S. Gill. A variational approach to two dimensional grain growth–I. Theory. *Acta materialia*, 44(12):4765–4775, 1996.

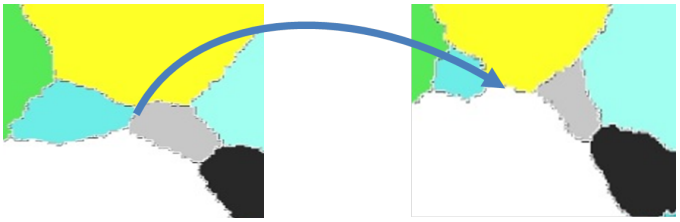
[3] N. R. Council et al. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. National Academies Press, 2008.

[4] S. Gill and A. Cocks. A variational approach to two dimensional grain growth–II. Numerical results. *Acta materialia*, 44(12):4777–4789, 1996.

[5] E. Payton. Revisiting sphere unfolding relationships for the stereological analysis of segmented digital microstructure images. *Journal of Minerals & Materials Characterization & Engineering*, 11(3):221–242, 2012.

[6] E. Payton, G. Wang, M. Mills, and Y. Wang. Effect of initial grain size on grain coarsening in the presence of an unstable population of pinning particles. *Acta Materialia*, 61(4):1316–1326, 2013.

(a) An example neighbor switch (T1 change)



(b) An example elimination of a grain (T2 change)

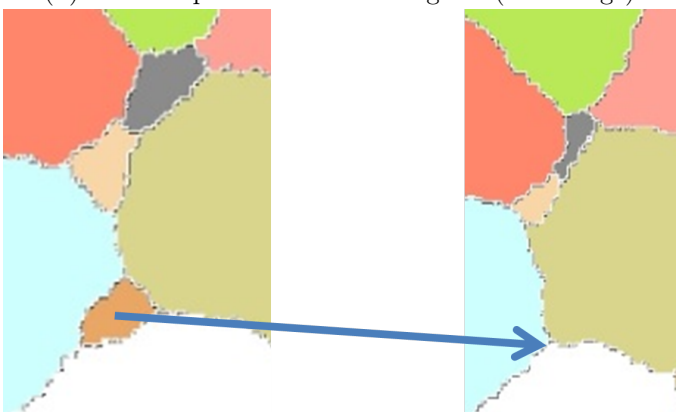


Figure 1: Examples of the two considered network topology changes from our simulations

[7] E. J. Payton. *Characterization and Modeling of Grain Coarsening in Powder Metallurgical Nickel-Based Superalloys*. PhD thesis, The Ohio State University, 2009.

[8] P. Rios and D. Zöllner. Grain growth–unresolved issues. *Materials Science and Technology*, pages 1–10, 2018.

[9] D. Zöllner. On the Aboav–Weaire-law for junction limited grain growth in two dimensions. *Computational Materials Science*, 79:759–762, 2013.

THE INFLUENCE OF TIME SERIES DISTANCE FUNCTIONS ON CLIMATE NETWORKS

Leonardo N. Ferreira, Liang Zhao, Elbert E. N. Macau

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Introduction

Complex networks has established itself as an important tool for the analysis of complex systems. In the context of climate systems, networks can be constructed (Fig. 1) using a spatiotemporal climate dataset and a time series distance function. It consists of representing each spatial area by a node and connecting nodes that present similar climate variations. The vast majority of papers in the literature [1] use the Pearson correlation coefficient or cross-correlation to measure the similarity between time series. In this paper, we study the influence of 29 time series distance functions on climate network construction using network theory.

Methodology

In our experiments, we constructed climate networks as described in Fig. 1. We used the NCEP/NCAR dataset [2] of monthly means records of air temperature near the surface and the world grid cells (2.5^o latitude $\times$ 2.5^o longitude) from 1948 to 2016. The resulting data is composed of 10512 time series with 816 values each. We removed the seasonal component using an additive decomposition by moving averages and normalized all time series between 0 and 1. Then, we calculate the distance between every pair of time series using 29 distance functions: Autocorrelation Coefficients (ACF), Avg(L_1, L_∞), Bhattacharyya, Compression-based, Complexity Invariant, Correlation, Cross-correlation, Dice, DISSIM, Dynamic Time Warping (DTW), Euclidean, Fourier Coefficient, Gower, Integrated Periodogram (INTPER), Jaccard, Kulczynski, Lorentzian, Manhattan, Mutual Information, Maximal Information Coefficient, Normalized Compression, Partial Autocorrelation Coefficients, Periodogram, Sorensen, Squared-Euclidean, Short Time Series, Tanimoto, Temporal Correlation and Raw Values, and Wave Hedges. Instead of choosing the same threshold value (τ) for every distance function, we use the p^{th} distance percentile as τ to build the networks with the same edge density.

Experimental Results and Analysis

According to our result, for all distance functions, there is an edge density interval between 0.1% and 1% where the networks become connected. Outside this range, the networks are very fragmented or too densely connected. Therefore, we consider this interval interesting for the study of climate networks. In this interval, almost all the networks present the small-world feature with degree distribution that decays sharper than a power-law distribution, what indicates that climate networks are not scale-free.

We observed that the distance functions used in the literature [2] generate similar networks. Interestingly, we verified that other distance functions create very different networks. These networks have differences in their long-distance connections (teleconnections). We define teleconnections as the most similar pairs of time series whose grid cells have great-circle distance longer than $5000km$. Short-distance connections were not considered because they tend to be similar since there is a spatial continuity of modeled phenomenon in the physical field of close nodes. In our experiments, we used the Haversine distance to compute the great-circle distance between two coordinates. We use the center coordinates of each grid cell to calculate the distance. We also limit the number of distance functions. Instead of studying the network topology generated by all the 29 distance functions, we limited ourselves to four ones.

Fig 2 illustrates the 500 strongest teleconnections for four different distance functions. As demonstrated by previous works [1], the correlation captures the influence of El Niño-Southern Oscillation (ENSO). ENSO is a very important teleconnection pattern responsible for climate variations all over the globe. This phenomenon is defined as a coupled ocean-atmosphere interaction in the Pacific. This interaction connects a large-scale oceanic sea surface temperature (SST) anomaly on the tropical Pacific (El Niño) to the large-scale atmospheric Southern Oscillation (SO), which is characterized by a sea-level pressure seesaw between French Polynesia and North Aus-

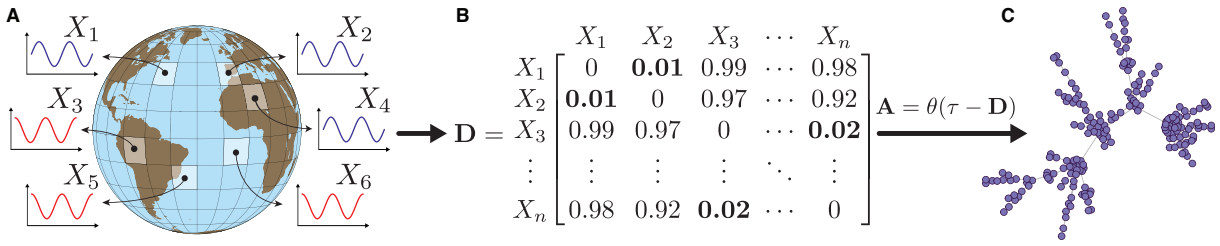


Figure 1: Climate network construction. (A) The first step consists of building a spatiotemporal climate dataset. This dataset can be constructed by dividing a region into grid cells that represents a climate variation of that smaller area, e.g., air temperature, sea pressure, precipitable water or wind speed. (B) Then, a time series distance function can be used to build a distance matrix $\mathbf{D}$ whose positions $\mathbf{D}_{ij}$ is the distance between two time series X_i and X_j; (C) The distance matrix $\mathbf{D}$ can be transformed into an adjacency matrix $\mathbf{A}$ by taking all the values from $\mathbf{D}$ lower than an arbitrary threshold value τ, i.e., $\mathbf{A} = \theta(\tau - \mathbf{D})$, where $\theta(\bullet)$ is the Heaviside step function. High threshold values ($\tau \approx 1$) generate highly connected networks while lower ones ($\tau \approx 0$) create disconnected networks.

tralia. This pattern is also captured by ACF, DTW and INTPER distances. The DTW distance also captures The Pacific-South American (PSA) pattern. The PSA modes are unique features of atmospheric variability in the Southern Hemisphere (SH). They represent a zonal wave with wavenumber 3 and a well-defined wave train from the tropical Pacific and Indian Ocean section to South America with large amplitude in the PSA sector.

Through INTPER, it is possible to recognize the relationship between the eastern Pacific and tropical Atlantic temperatures that can be explained by the Southern Oscillation. The ACF and INTPER show an interesting connection between Indian, Pacific and Atlantic oceans. Although there is no teleconnection pattern that explains this relationship, studies have already shown that the temperatures of this three basins are interconnected. Those results show that different distance functions are capable to capture different patterns and relationships that are not very well understood such as those shown in ACF and INTPER.

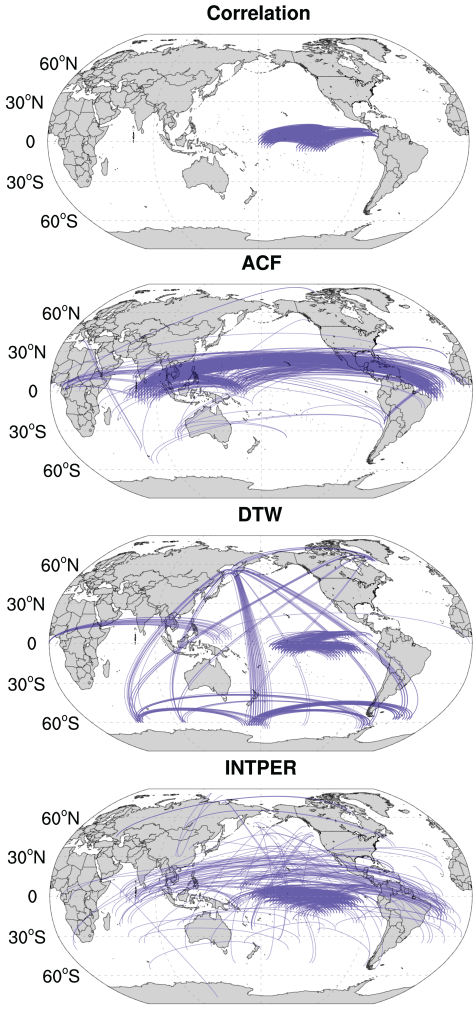


Figure 2: Teleconnections for four distance functions: Correlation, ACF, DTW and INTPER. We considered only the 500 strongest teleconnections (nodes distance grater than $5000km$).

Conclusions

In this paper, we analyzed the influence of 29 different time series distance functions on the topology of climate networks using network theory. We observed differences in the topologies of climate networks generated by different distance functions. These distance functions capture information that previous analysis have not observed. These patterns represent an improvement to atmospheric sciences knowledge and should be further investigated.

References

[1] J. F. Donges, Y. Zou, N. Marwan, and J. Kurths. Complex networks in climate dynamics. *The European Physical Journal Special Topics*, 174(1):157–179, Jul 2009.

[2] E. Kalnay et al. The ncep/ncar 40-year reanalysis project. *Bulletin of the American Meteorological Society*, 77(3):437–471, 2015/10/05 1996.

NEURAL INTERACTOME: INTERACTIVE SIMULATION OF A NEURONAL SYSTEM

Jimin Kim, William Leahy, Eli Shlizerman

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Emerging neuro-imaging techniques and novel optical interfaces which record and control neural dynamics enable detailed computational connectivity and dynamics models for neuro-biological systems. An open question stemming from these advances is how to validate, simulate and apply these models to predict network functionality. We therefore develop a platform to inspect network dynamics in real time while preserving structural connectivity properties, displaying the dynamics on a graph, with possibilities to identify functional sub circuits and review the simulated dynamics.

Abstract

Both connectivity and biophysical processes determine the functionality of neuronal networks. We, therefore, develop a real-time framework, called Neural Interactome, to simultaneously visualize and interact with the structure and dynamics of such networks. Neural Interactome is a cross-platform framework, which combines graph visualization with the simulation of neural dynamics, or experimentally recorded multi neural time series, to allow application of stimuli to neurons to examine network responses. In addition, Neural Interactome supports structural changes, such as disconnection of neurons from the network (ablation feature), as typically done in experiments. Neural dynamics can be explored on a single neuron level (using a zoom feature), back in time (using a review feature) and recorded (using presets feature). We implement the framework using a model of the nervous system of Caenorhabditis elegans (C. elegans) nematode, a model organism for which full connectome and neural dynamics have been resolved. We show that Neural Interactome assists in studying neural response patterns associated with locomotion and other stimuli. In particular, we demonstrate how stimulation and ablation help in identifying neurons that shape particular dynamics. We examine scenarios that were experimentally studied, such as touch response circuit, and explore new scenarios that did not undergo elaborate experimental studies. The development of the Neural Interactome was guided by generic concepts to be applicable to neuronal networks with different neural connectivity and dynamics.

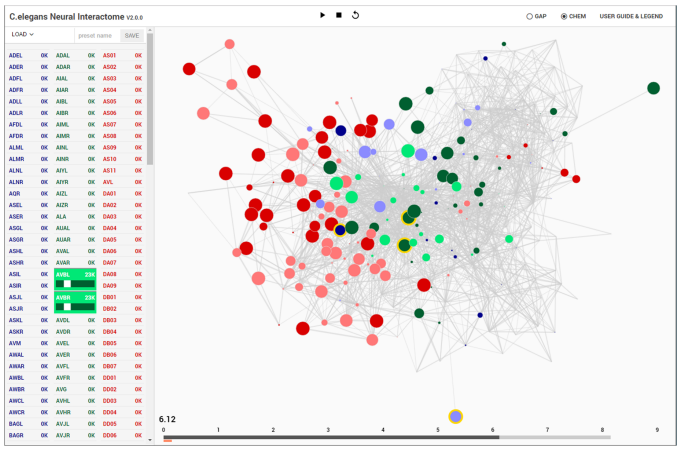


Figure 1: **Interactive Interface of Neural Interactome**. Left panel enlists all the neurons classified by type (sensory, inter and motor). Each neuron is a clickable button with a scroll option. Scrolling adjusts the magnitude of constant stimulus; click+shift ablates the neuron from the network. Right: Force-directed graph displays each neurons membrane voltage (node color denotes the sign and radius denotes the magnitude) and connections between neurons (edges between each pair of nodes). At the bottom of the graph, time bar keeps track of visualization current time (dark gray), and of computed time by the backend neural integration (light gray).

APPROXIMATING TRACE($L^\dagger$) USING THE JOHNSON-LINDENSTRAUSS LEMMA AND SPECTRAL METHODS.

Jesse Laeuchli

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We introduce a method for finding the Kirchhoff index of very large graphs using the Johnson-Lindenstrauss lemma and Spectral Methods.

Introduction

Computing the trace of the inverse of $L^\dagger$ is a problem that occurs in many areas such as Lattice Quantum Chromodynamics, Machine Learning, Uncertainty Quantification, Mathematical Chemistry and Network Analysis [2], [1]. In several of these areas, the matrix L in question is a graph Laplacian, and the trace of the pseudo-inverse of this graph is needed. An example of this would be to obtain the Kirchhoff index n trace($L^\dagger$).

While exactly solving for trace($L^\dagger$) is infeasible when the matrix is large, there are several statistical methods for estimating trace($L^\dagger$). A common one is Hutchinson's method [3]. In this algorithm, several vectors are prepared with their elements being drawn from the Rademacher distribution. The trace estimate is then $z'L^\dagger z$, where the matrix-vector multiplication is computed using a linear solver. This method can obtain a relatively low accuracy cheaply, but to improve beyond this takes a significant number of additional vectors. This can be expensive since each matrix-vector operation requires the invocation of a linear solver.

Statistical methods such as Hutchison's are the best that can be done if the matrix lacks structure, meaning that no elements of the inverse are much larger or smaller than the average element. However, if the matrix has such a structure and it can be effectively discovered then probing vectors can be created which are more effective than statistical approaches. This is done by effectively dropping or ignoring the smallest elements of the inverse and coloring the rest. By grouping the remaining nodes of the same colors together a block zero structure is induced along the diagonal and the probing vectors can then be used to recover the trace, albeit with the error introduced by treating certain elements of the inverse as zero when they are not. Probing methods can be quite effective and can still provide statistical bounds on the error of the resultant trace. If the relative size of the elements can be modeled with a linear function, then the variance of a trace estimate found by probing can decrease as $\frac{1}{n}$ (where n is the number of vectors needed) instead of the $\frac{1}{\sqrt{n}}$ of the statistical methods, a substantial improvement.

A major problem with probing based methods is that it can be difficult to discover the structure of $L^\dagger$. One approach is to take powers of L and color this as an approximation for the important elements of $L^\dagger$. If the Neumann series converges to $L^\dagger$ with high accuracy in a limited number of steps then this can work well. However, for many graphs of interest such as those of real-world social networks this method works very poorly and probing does not surpass statistical methods. Alternatively, one can use spectral methods. By finding the smallest eigenvector of $L^\dagger$ one can find an approximate red-black coloring that can be used to create probing vectors. By then recursively finding the smallest eigenvector of the red-black sub-blocks, the probing vectors can be improved. Unfortunately finding the eigenvectors of these sub-blocks of $L^\dagger$ is very expensive since each matrix-vector multiplication with $L^\dagger$ requires a linear solve.

Our Proposal

We propose a new method that leverages both statistical and structural approaches. By using the Johnson-Lindenstrauus lemma we obtain an approximation to $L^\dagger$, by projecting the elements of $L^\dagger$, into a lower dimensional space. This is the same approach as was proposed by Spielman and Srivastava [4] for effectively calculating the effective resistances given a fast solver, and can be done using a relatively low number of linear solves at a low accuracy. While we could easily use this approximation to the elements of $L^\dagger$ to compute the trace, the error we would obtain is the same as with Hutchinson's method. However, this gives us a way to cheaply obtain an approximation to the eigenvectors of the sub-blocks of $L^\dagger$, since

we no longer require a linear solve for each matrix-vector operation. While the eigenvectors will not be computed very accurately, this will not greatly impact the quality of the probing vectors, because with structural methods we care about the relative size of the elements, and not their absolute value. Unlike other structural approaches, this method is relatively easy to analyze because the error introduced by the Johnson-Lindenstrauus lemma is known, and the Cheeger bounds provide grantees about the quality of the given coloring. It also provides a good way to analyze the quality of our probing vectors, because the spectral gap provides an indication of the quality of the partitions our algorithm is producing.

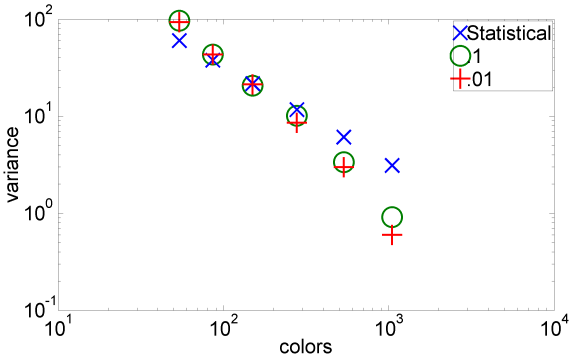


Figure 1: Example of our method being used to estimate the trace of a random scale free graph. The results show a comparison of the statistical method against our method, where the vectors used to build the Johnson-Lindenstrauus projection are solved with two different low tolerances.

We have tested this method against several social media graphs arising in the real world, as well as several synthetic graphs which are thought to model real networks well and have obtained better accuracy than both pure statistical and structural methods using the same number of linear solves.

References

E. Estrada, N. Hatano, *Statistical-mechanical approach to subgraph centrality in complex networks*, Chemical Physics Letters, 439, (5/07), pp. 247-251.

A. Stathopoulos, J. Laeuchli, and K. Orginos, *Hierarchical probing for estimating the trace of the matrix inverse on toroidal lattices* , SIAM J. Sci. Comput.,35(5) (2013), pp. 299–322.

H. Avron and S. Toledo, *Randomized algorithms for estimating the trace of an implicit symmetric positive semi-denite matrix*, Journal of the ACM, 58 (2011), p. Article 8.

Daniel A. Spielman and Nikhil Srivastava,*Graph Sparsification by Effective Resistances* , CoRR, abs/0803.0929, (2008)

AN ADAPTIVE MULTICAST SCHEME IN MOBILE AD HOC NETWORKS

Hyunsun Lee, Yi Zhu, Brian Spain

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Abstract

Ad hoc networks are a crucial foundation for Machine-to-Machine (M2M) communications that can be disconnected, intermittent, and limited. Such limitations have been addressed and researched in literature. In this study, we propose an adaptive multicast scheme (Smart Multicast) for Mobile Ad Hoc Networks (MANETs), based on a modified stochastic branching process, to support data exchange in fast switching topology. Our approach aims at balancing reachability, total node usage, and average branching factor in multi-hop data dissemination by locally regulating the transmission probability and adaptively selecting neighbor nodes.

The stochastic branching process model is often used to describe the beginning stage of a disease outbreak when the number of infected patients remains significantly smaller than the entire population. The network of individual contacts and transmissions are focused in the branching process model. The importance of the method is placed on the transmission rate from one individual to the next generation, that guides us to understand the overall structural pattern of the transmission. In the stochastic branching model, the basic reproductive number, R, is a significant indicator to classify whether an infectious diseases is a minor outbreak or a major outbreak. This reproductive number is modified and used to determine the local transmission probability in the proposed model.

We first begin with our problem setting of network. We consider a finite rectangle domain in two-dimensional space. There are fixed number of moving objects (nodes) with the same speed and transmission range in the domain. Within the fixed transmission range, the 100% connection is assumed to be guaranteed. Each node is moving in a random direction and bounces back inside the domain when it reaches the boundary of the domain, so that the number of moving nodes stays constant during the simulation. There is a source node and a destination node in the pool of moving nodes where the source node in the 0-th generation initiates the transmission by broadcasting a message to all neighbor nodes within the transmission range. If the destination is within the transmission range of a node then the communication is terminated and it is marked as a success.

The major difficulty in adapting the stochastic branching process to the wireless sensor network is that we do not have access to the entire picture of the network, rather the network constantly changes in time as the nodes moves in the domain. To obtain the most information of the fast switching network, we adapt a statistic along each transmission path and the statistic is used to understand the structure of the transmission path so that we can dynamically control the transmission rate at each node. The statistic is calculated similarly to the reproductive number R_n as in [1] along the transmission path at the current node in the n-th generation. We also consider the number of neighbors, d_n, within the transmission range that have not received or transmitted the message in any previous transmissions. The transmission probability T_n at the current node is adjusted from the previous transmission probability as well as the local densities as follows:

$$T_n = T_{n-1} + T_{n-1} \cdot \omega(d_{n-1}, d_n, R_{n-1}, R_n), \qquad (1)$$

where ω is a weight function depending on d_{n-1}, d_n, R_{n-1}, and R_n. This weight function ω can be determined depending on the network situation. In our current setting, we set the weight function as:

$$\omega(d_{n-1}, d_n, R_{n-1}, R_n) = 1 - \frac{R_n/d_n}{R_{n-1}/d_{n-1}}. \qquad (2)$$

The transmission probability T_n should be between 0 and 1. If the calculated T_n in Equation (1) is below 0, then we set it to $1/d_n$ and the message will be sent to only one available neighbor node. If it is above 1, then we set it to 1 and the message will be broadcasted to all neighbor nodes within the transmission range that never received the message in any previous transmissions, called active neighbor nodes. The distribution probability in Equation (1) increases if the current density in the n-th generation is larger than the density at the previous node

in the $(n-1)$-th generation, meaning there is a better chance to disseminate the message effectively when there are relatively more neighbors. However, the probability is controlled based on the reproductive numbers, that is, if the current statistical reproductive number is greater than that at the previous node, it suppresses the distribution at the current node.

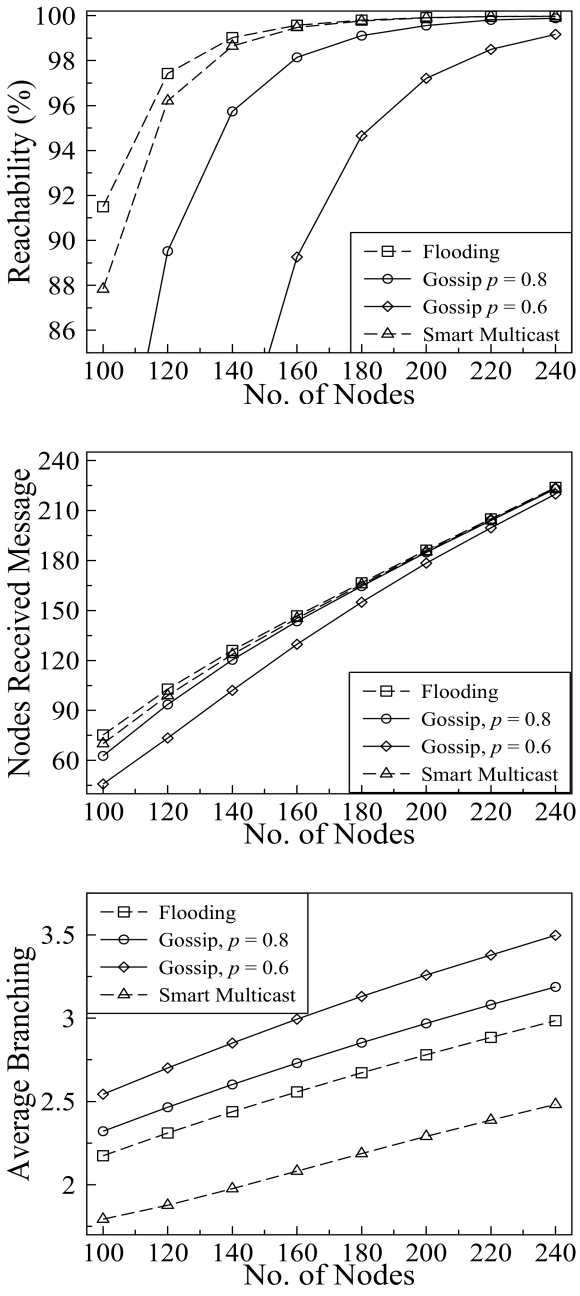


Figure 1: Numerical result comparison: reachability (top), total node usage(middle) and average branching factor (bottom).

When the message is transmitted to $T_n \cdot d_n$ active neighbors from the current node, the priority of choosing the neighbors is determined by the distance between the current node and the neighbor node, or by the negative velocity correlation. For the distance based selection, we select the active neighbor nodes from the farthest ones for transmission so that the message propagates faster and possibly covers more area in its vicinity. For the velocity based selection, the active neighbor node that moves relatively fastest away from the current node has the highest priority. One may also set the priority differently depending on any required energy saving mode. In the poster, we present the numerical results using the distance based selection.

The performance of our proposed algorithm is evaluated and compared with pre-existing well-known algorithms such as flooding and gossiping algorithms in [2, 3, 4] in the poster presentation. The comparison is made based on the reachability, the total node usage, and average branching factor as in Figure 1. The reachability is calculated by the rate (%) of successes based on 100,000 simulations, the total node usage is the number of nodes that received and sent the message during the transmission, and the average branching factor is an averaged ratio between the total number of transmitted messages and the total number of sending nodes. Our proposed model achieves high reachability close (but slightly less) to flooding method, the comparable node usage to the gossip method, and outstanding average branching factor among the compared methods.

The major contribution of our work is to adopt the reproductive number, R_n, to acquire the statistical information of the network structure along the transmission path and the local density, d_n, to dynamically control the transmission probability locally. The detailed algorithm and explanation will be provided in the presentation.

References

[1] F. Braucer and C. Castillo-Chavez. *Mathematical Models in Population*. Springer Biology and Epidemiology, 2011.

[2] Z. J. Haas, J. Y. Halpern, and L. Li. Gossip-based ad-hoc routing. *IEEE/ACM Transactions on Networking*, 14(3):479–491, 2006.

[3] A. Rahman, W. Olesinski, and P. Gburzynski. Contentrolled flooding in wireless ad-hoc networks. *Proc. International Workshop on Wireless Ad-Hoc*, 2004.

[4] D. Shah. Gossip algorithms. *Foundations and Trends in Networking*, 3(1), 2008.

HYPOTHESIS TESTING FOR DYNAMIC NETWORKS

Nathan Lemons

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Motivated by the problem of detecting anomalous subgraphs in a graph of computer network connections we consider the following problem: given a time series observation of a graph $G = (G_1, G_2, \ldots, G_m)$, when can we determine if G was generated under the temporal graph null model H_0 or under the alternative H_1? This is a wide open problem in network inference with applications in control systems, cyber-security, and neuroscience.

We consider simple Markovian random graph models where the graph at time t depends only on the graph at time $t-1$. We investigate and compare various tests, from both an algorithmic and statistical point of view, to determine when the models are distinguishable.

Introduction

Dynamic networks appear in a great variety of contexts such as spreading of infectious disease [12], synchronization of electric power generators [5], learning in the brain [3], and computer communication systems [1]. While there is an increasingly large interest and literature in dynamic graphs, there has been relatively little attention paid to dynamic graphs from the statistical community. In contrast, there is a large literature regarding hypothesis testing in the static graph case (see [11] and the references therein).

Motivated by publicly available data [8], including labeled attacks [10, 9], collected from authentication events on the Los Alamos computer network, we construct hypothesis tests for temporal graph models using methods and ideas from [2]. In the motivating data sets, global graph statistics such as edge counts; vertex degrees; and small motif counts are stable, while approximately 25% of edges change from day to day [6].

Models

We build our hypothesis tests around the following Markovian model of temporal graphs [4, 7].

Definition 1 (Edge Markovian)**.** A random temporal graph $\{G_t\}$ on n vertices is called *edge Markovian* if there exist maps $P, Q : E(K_n) \to [0,1]$ such that for all $t \geq 0$, each edge e of G_{t+1} is determined independently with probability

$$\mathbb{P}[e \in G_{t+1}] = \begin{cases} P(e) \text{ if } e \notin G_t\,, \\ 1 - Q(e) \text{ if } e \in G_t\,. \end{cases} \tag{1}$$

We consider the problem of distinguishing between two different edge Markovian graph distributions. In the null model, all vertices are homogeneous, while in the alternative model a subset of nodes will have larger degrees.

Definition 2 (Null Hypothesis)**.** Fix $\alpha > 0$. The graphs $\{G_t\}$ are edge Markovian with G_0 distributed as an Erdős-Rényi random graph $G(n, p_0)$ and for all edges e,

$$P(e) \equiv \alpha p_0\,, \tag{2}$$
$$Q(e) \equiv \alpha(1 - p_0)\,. \tag{3}$$

In the alternative model there exists a subset of vertices which are more likely to appear in edges.

Definition 3 (Alternative Hypothesis)**.** Let α and p_0 be as in the Null Hypothesis. Fix $p_1 > p_0$ and an integer k. The graphs $\{G_t\}$ are edge Markovian with G_0 distributed according to $G(n, p_0)$ and there exists a subset S of k vertices such that P and Q are given by

$$P(e) = \begin{cases} \alpha p_0 \text{ if } e \text{ is not incident to } S\,, \\ \alpha p_1 \text{ if } e \text{ is incident to } S\,, \end{cases} \tag{4}$$

$$Q(e) = \begin{cases} \alpha(1 - p_0) \text{ if } e \text{ is not incident to } S\,, \\ \alpha(1 - p_1) \text{ if } e \text{ is incident to } S\,. \end{cases} \tag{5}$$

Note that in the alternative hypothesis, edges in G_t incident to S occur with probability $(1-\alpha)^t p_0 + [1 - (1-\alpha)^t]\, p_1$; all other edges occur with probability p_0. The testing problem can be generalized to the setting in which the change occurs at some unknown time l, though we do not consider that here.

Results

A test function, f, is a binary function defined on the input $\mathbf{X} = \{G_t\}$. We consider the risk of a such a function

measured by

$$R(f) = \mathbb{P}_0\left[f(\mathbf{X}) = 1\right] + \frac{1}{\binom{n}{k}} \sum_S \mathbb{P}_S\left[f(\mathbf{X}) = 0\right] . \quad (6)$$

We prove the following.

Theorem 1. *If*

$$|S|(p_1 - p_0)\sqrt{\frac{t\alpha}{p_1}} \to \infty,$$

then there exists a constant M such that $R(f)$ approaches 0 when f is the total degree-based test:

$$f(\mathbf{X}) = \begin{cases} 0 & \text{if } \sum_{i=0}^{t} e(G_i) < M\,, \\ 1 & \text{otherwise.} \end{cases} \quad (7)$$

For static graphs, there are regions of parameter space where the total degree test cannot distinguish the two models, but in which other tests can, such as

- the scan test,
- size of the largest connected component test, and
- the subtree test

Question 2. Are there variants of these tests which are algorithmically fast *and* work for temporal graphs?

References

[1] N. Adams and N. Heard, editors. *Data Analysis for Network Cyber-Security*. World Scientific, 2014.

[2] L. Addario-Berry, N. Broutin, L. Devroye, G. Lugosi, et al. On combinatorial testing problems. *The Annals of Statistics*, 38(5):3063–3092, 2010.

[3] D. S. Bassett, N. F. Wymbs, M. A. Porter, P. J. Mucha, J. M. Carlson, and S. T. Grafton. Dynamic reconfiguration of human brain networks during learning. *Proceedings of the National Academy of Sciences*, 108(18):7641–7646, 2011.

[4] A. E. Clementi, C. Macci, A. Monti, F. Pasquale, and R. Silvestri. Flooding time of edge-Markovian evolving graphs. *SIAM journal on discrete mathematics*, 24(4):1694–1712, 2010.

[5] F. Dörfler, M. Chertkov, and F. Bullo. Synchronization in complex oscillator networks and smart grids. *Proceedings of the National Academy of Sciences*, 110(6):2005–2010, 2013.

[6] A. Hagberg, N. Lemons, A. Kent, and J. Neil. Connected components and credential hopping in authentication graphs. In *SITIS 2014*, pages 416–423. IEEE, Nov 2014.

[7] A. Hagberg, N. Lemons, and S. Misra. *Dynamic Networks and Cyber-Security*, volume 1 of *Security Science and Technology*, chapter Temporal reachability in dynamic networks. World Scientific, May 2016.

[8] A. D. Kent. Anonymized user-computer authentication associations in time, 2014.

[9] A. D. Kent. Comprehensive, Multi-Source Cyber-Security Events. Los Alamos National Laboratory, 2015.

[10] A. D. Kent. Cybersecurity Data Sources for Dynamic Network Research. In *Dynamic Networks in Cybersecurity*. Imperial College Press, June 2015.

[11] G. Lugosi. Lectures on combinatorial statistics. 2017.

[12] M. E. J. Newman. Spread of epidemic disease on networks. *Phys. Rev. E*, 66:016128, Jul 2002.

BIASED EDGE-WEIGHTING SCHEMES CAN BOOST REPRODUCIBILITY OF CENTRALITY MEASURES IN BRAIN NETWORKS

Tingshan Liu, Gwen Spencer

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Graph-theoretic measures in brain networks are potentially capable of identifying biomarkers. However, if the discrepancy among groups (conditions) is not stable under re-sampling, the biomarkers may be due to variability instead of systematic differences. In this work, we propose several network construction schemes to promote the poor reproducibility of centrality measures in the traditional cortical thickness brain networks.

Introduction

Centrality measures have been widely used to study functional and structural brain networks, and are often related to information integration, network efficiency and differentiation of neurological diseases [4][3].

Consider a pair of brain networks from two populations (conditions). We may want to compare the two sets of k regions with the highest centrality in these networks in order to study the overlapping and non-overlapping regions if our goal is to evaluate the 'importance' of a certain region with respect to this particular condition. One question we are interested in is how stable this overlap is. In other words, how reliable are these biomarkers generated from the group discrepancies? More importantly, how can we improve the reproducibility? If this pair of networks represent two subsets of the same population, but the size of the overlap is much smaller than k using the traditional network construction method, we would probably expect a fairly large false positive rate in the study. In this work, we evaluate the reproducibility of two networks generated from the same population and manipulate the weighting scheme in order to promote reproducibility.

Evaluating Reproducibility

In order to quantify the stability of a given query, Q, we measure the size of the overlap between two answers to Q obtained from two random pairs of networks. We focus on the type of queries, Q_k: Of the brain regions considered, which k are most central with respect to a certain centrality measure? The *reproducibility* of Q_k for sample size s, the number of sets of cortical thickness data included, is defined by

$$R^{\mathcal{N}}(Q_k, s) = E(|O_{Q_k}|),$$

the expected value of the size of overlap where $\mathcal{N}$ denotes the brain network construction method. For a given combination of $\mathcal{N}$, Q_k and s, we perform 3,000 random pairs of consecutive experiments to approximate $R^{\mathcal{N}}(Q_k, s)$. Another type of queries, Q_{sp}, consider the entire network of n nodes, instead of the *top-k-hubs* as Q_k. For a pair of experiments, Exp1 and Exp2, we calculate Spearman's footrule distance as follows, Spearman(Exp1, Exp2) $= \sum_1^n |X(i) - Y(i)|$, where $X(i)$ and $Y(i)$ denote the rank of a region (node) i specified by Exp 1 and Exp 2, respectively.

Network Construction

We obtained 298 sets of cortical thickness data from 118 participants who showed no symptoms in the Clinical Dementia Rating. For each sample size s and for each pair of regions, Pearson correlation coefficients were calculated across all s participants. Apart from constructing binary graphs based on p-values, as implemented in various studies [2][1], we also conduct our biased edge-weighting schemes to build weighted graphs of 68 nodes (regions).

- Normal Ranked Weighting Scheme (correlation coefficient (r)-based or p-based): Based on the ranking of either p or r values, an index between 1 to $\binom{68}{2} = 2278$ was assigned to each edge.
- Ordinal Ranked Weighting Scheme (r-based or p-based): Based on the normal ranking of p or r values, a set of 2278 ranked edges was divided into 67 equivalent classes such that each equivalent calss E_i contains i edges.

Results

Using consistent sample sizes for different condition groups is not required in the construction of cortical thickness cor-

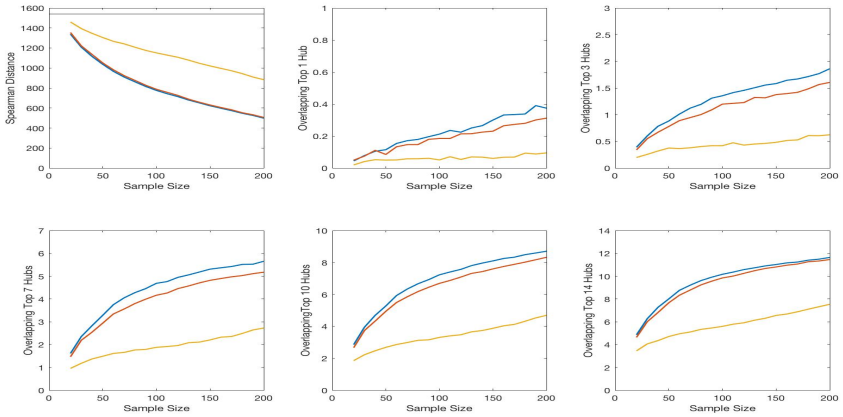


Figure 1: Spearman's Distance and Overlapping top-k-hubs queries for degree centrality (—), closeness centrality (—) and betweenness centrality (—) using 10% p-binary graph.

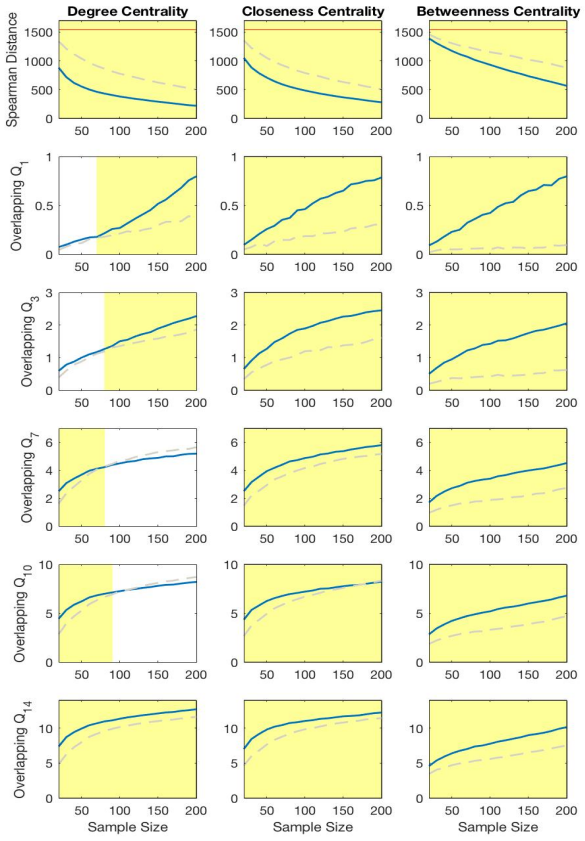


Figure 2: Ordinal r-based edge-weighting strongly boosts reproducibility for closeness and betweenness centrality. Regions where our alternative technique is dominant are highlighted in yellow.

relation networks. However, binary p-value-based network construction has very poor reproducibility for centrality queries with small sample sizes (Fig. 1).

In order to evaluate the reproducibility improvement for each weighting scheme, we introduce the Boosting Factor (BF),

$$BF^{\mathcal{N}}(Q_i, s) = \frac{\min_{s'}\{s' : R\text{ p-binary}(Q_i, s') \geq R^{\mathcal{N}}(Q_i, s)\}}{s}$$

For example, $BF^{\mathcal{N}}(Q_7, 50) = 150\%$ denotes that switching to method $\mathcal{N}$ to construct a network of 50 samples is comparable to including 75 samples in the traditional network in terms of reproducibility. Hence, a large BF suggests the potential of a network construction scheme to achieve the same reproducibility but with much smaller cost. Our method is particularly effective for closeness and betweeness centrality measures. We obtained $BF = 150 - 200\%$ for most Q_i's (Fig. 2).

References

[1] S. Achard, R. Salvador, B. Whitcher, J. Suckling, and E. Bullmore. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. *Journal of Neuroscience*, 26(1):63–72, 2006.

[2] Y. He, Z. J. Chen, and A. C. Evans. Small-world anatomical networks in the human brain revealed by cortical thickness from mri. *Cerebral cortex*, 17(10):2407–2419, 2007.

[3] A. Özgür, T. Vu, G. Erkan, and D. R. Radev. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285, 2008.

[4] M. P. van den Heuvel, C. J. Stam, R. S. Kahn, and H. E. H. Pol. Efficiency of functional brain networks and intellectual performance. *Journal of Neuroscience*, 29(23):7619–7624, 2009.

NORMALIZED MUTUAL INFORMATION EXAGGERATES COMMUNITY DETECTION PERFORMANCE

Arya D. McCarthy, David W. Matula

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

We present a critical evaluation of normalized mutual information (NMI) as an evaluation metric for community detection (CD). NMI exaggerates the leximin method's performance on weak communities: Does leximin, in finding the trivial singletons clustering, truly outperform eight other CD methods? Three NMI improvements from the literature are AMI, rrNMI, and cNMI. We show equivalences under relevant randomness models, and **for CD evaluation, we advise one-sided AMI under** $\mathbb{M}_{\text{all}}$ (all partitions of n nodes). This work seeks 1) to start a conversation on robust measurements, and 2) to advocate evaluations which do not give "free lunch".

Background

Unsupervised algorithms—like those for CD—present a challenge for appraisal. In CD, we circumvent this by using external evaluation tasks. Practitioners apply CD methods to benchmark graphs containing "ground truth" communities, then compute an agreement measure to determine how well those communities are recovered.

The popular measure in CD is NMI. Its theoretical flaws have been noted [4, 7, 9, 11]. Particularly relevant is the non-homogeneity of the measure: NMI awards credit for low-information guessing [7]. This deficiency has demonstrable implications for method selection, which we later show using the leximin method.

Improvements to NMI have gained little traction in CD. A sequence of proposed improvements in the CD community led to the recent *corrected NMI* (cNMI) [4]. An older measure common outside of CD, *adjusted mutual information* (AMI) [9], has been analyzed and advocated for CD [6, 7]. AMI augments NMI's consistent upper bound (1.0) with a consistent zero expectation by adjusting for chance clusterings:[1]

$$\text{AMI}_{sqrt}(\hat{C}, C) = \frac{I(\hat{C}, C) - \mathbb{E}_{\hat{C}', C'}\left[I(\hat{C}', C')\right]}{\sqrt{H(\hat{C}) \cdot H(C)} - \mathbb{E}_{\hat{C}', C'}\left[I(\hat{C}', C')\right]}.$$

[1] Negative AMI indicates worse-than-chance clusterings.

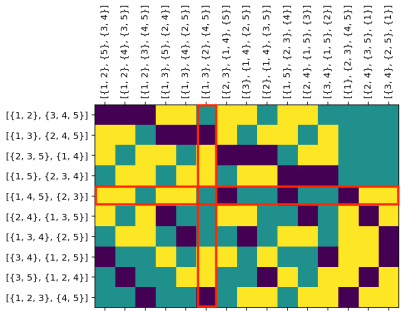


Figure 1: AMI computes expectation over all (X', Y') pairs in the joint distribution. cNMI examines only the observation's row and column. In $\mathbb{M}_{\text{perm}}$ (shown), exchangeability under MI makes these expectations equal.

Many forms of NMI and derived measures exist. We color-code the averaging method in magenta. Including the expectation model, in blue, differentiates AMI from NMI.

Excellent recent work [3] extends AMI to other models of randomness. The literature has implicitly computed expectations over $\mathbb{M}_{\text{perm}}$: all partitions of the same *class* (or cluster-size pattern) as the observation. More correct for CD is an expectation over $\mathbb{M}_{\text{all}}$: all partitions of n nodes. The work also devises one-sided randomness models for comparing against a fixed ground truth.

Theoretical Results

The cNMI between two partitions C and $\hat{C}$ is:

$$\frac{2\text{NMI}(\hat{C}, C) - \left(\mathbb{E}_{\hat{C}'}\left[\text{NMI}(\hat{C}', C)\right] + \mathbb{E}_{C'}\left[\text{NMI}(\hat{C}, C')\right]\right)}{2 - \left(\mathbb{E}_{\hat{C}'}\left[\text{NMI}(\hat{C}', \hat{C})\right] + \mathbb{E}_{C'}\left[\text{NMI}(C, C')\right]\right)}.$$

Note the different arguments in the top and bottom expectations. Lai and Nardini created cNMI to symmetrize the asymmetric *ratio of relative NMI* [11]:

$$\text{rrNMI}(C, \hat{C}) = \frac{\text{NMI}(C, \hat{C}) - \mathbb{E}_{\hat{C}'}\left[\text{NMI}(C, \hat{C}')\right]}{\text{NMI}(C, C) - \mathbb{E}_{C'}\left[\text{NMI}(C, C')\right]}.$$

Regardless of the randomness model, holding one clustering fixed in cNMI gives rrNMI, so rrNMI equals the one-sided cNMI. We note that cNMI is the *mediant*, or

"freshman sum" of rrNMI$(C, \hat{C})$ and rrNMI$(\hat{C}, C)$ in all randomness models.

In $\mathbb{M}_{\text{all}}$, rrNMI also equals the 1-sided AMI. The equality is a straightforward result of holding one clustering fixed and the fact that the normalizing constant in $\mathbb{M}_{\text{all}}$ must be $\log n$ [3]. Still, AMI is not the mediant because a different expectation is computed.

Additionally, as Figure 1 shows, the expectation in $\mathbb{M}_{\text{perm}}$ when varying both clusterings equals the expectation when varying either cluster, so the numerators in AMI and cNMI are equal. Any two-sided adjusted measure is the mediant of its associated one-sided measures.

AMI and cNMI are irreconcilable in general, despite their one-sided equality in $\mathbb{M}_{\text{all}}$. In an $\mathbb{M}$ which makes the numerators equal, the denominators differ, and vice versa.

Finally, we recognize that four variants of NMI in $\mathbb{M}_{\text{perm}}$ and $\mathbb{M}_{\text{num}}$ are instances of one underlying function under various p-norms. Derived measures like AMI are also parameterizable. Each uses some generalized, or Hölder, mean M_p to compute its normalizing constant: sqrt for M_0, sum for M_1, max for M_∞, etc.—our magenta color-coding. Note that in $\mathbb{M}_{\text{all}}$, we take $M_p(\log n, \log n)$—choice of p-norm does not alter the normalizing constant.

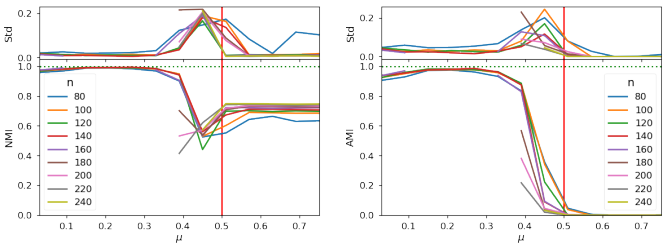
Experimental Results

We benchmark the AMI and NMI of eight popular CD methods, plus an adversarial case: the *leximin* method [5, 6], whose clustering is the dual of its *lexi*cographic max*imin* flow allocation [8]. We follow Yang et al. [10] in testing 25 LFR realizations each for various parameters, shown in Figure 2. With NMI, their method selection rules strongly prefer leximin for networks with weak communities. The adjusted-for-chance measure reveals that the recommendation is faulty.[2] We review the recommendations of Yang et al. and show that leximin, like many, is appropriate only for networks with strong communities.

Discussion

We see that the leximin method, like many, is successful on strong communities and is incorrectly appraised by NMI. Future work will assess CD methods and the related k-way partition problem using one-sided AMI under $\mathbb{M}_{\text{all}}$ and $\mathbb{M}_{\text{num}}$ respectively, as well as varying p. While NMI was the best known measure when Danon et al. first used it for CD [1], we see that **NMI can exaggerate community detection performance**. The choice of AMI corrects this exaggeration, so we encourage AMI's use in the CD community moving forward.

[2] While detectability limits are well-known for the stochastic block model [2], they are not in the LFR model. The performance of other methods shows that leximin falters before any detectability limit.



(a) Leximin method, NMI (b) Leximin method, AMI

Figure 2: Mean and st. dev. as a function of the mixing coefficient μ. Colors represent network sizes. NMI suggests that leximin succeeds on graphs with weak communities.

References

[1] L. Danon, A. Diaz-Guilera, J. Duch, and A. Arenas. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(09):P09008, 2005.

[2] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Physical Review E*, 84(6):066106, 2011.

[3] A. J. Gates and Y.-Y. Ahn. The impact of random models on clustering similarity. *The Journal of Machine Learning Research*, 18(1):3049–3076, 2017.

[4] D. Lai and C. Nardini. A corrected normalized mutual information for performance evaluation of community detection. *Journal of Statistical Mechanics: Theory and Experiment*, 2016(9):093403, 2016.

[5] C. F. Mann. *Extensions of maximum concurrent flow to identify hierarchical community structure and hubs in networks*. PhD thesis, Southern Methodist University, 2008.

[6] A. D. McCarthy. Gridlock in networks: The leximin method for hierarchical community detection. Master's thesis, Southern Methodist University, 2017.

[7] L. Peel, D. B. Larremore, and A. Clauset. The ground truth about metadata and community detection in networks. *Science Advances*, 3(5), 2017.

[8] F. Shahrokhi and D. W. Matula. The maximum concurrent flow problem. *Journal of the ACM (JACM)*, 37(2):318–334, 1990.

[9] N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct):2837–2854, 2010.

[10] Z. Yang, R. Algesheimer, and C. J. Tessone. A comparative analysis of community detection algorithms on artificial networks. *Scientific Reports*, 6:30750, 2016.

[11] J. Zhang, T. Chen, and J. Hu. On the relationship between Gaussian stochastic blockmodels and label propagation algorithms. *Journal of Statistical Mechanics: Theory and Experiment*, 2015(3):P03009, 2015.

EON: AN OPEN-SOURCE PYTHON PACKAGE FOR EPIDEMICS ON NETWORKS

Joel C. Miller

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

The modeling of epidemics on networks has grown over the last two decades, and there is now a need for an accessible tool allowing researchers to perform efficient simulations and implement standard analytic models. A new software package, EoN, implements many of the standard analytic models of SIS and SIR spread and provides efficient techniques for simulating both Markovian and non-Markovian dynamics. We describe the package, and because the simulation technique is faster and more flexible than Gillespie simulation, we describe it in more detail.

Please note, I would prefer that this be a poster.

Introduction

Epidemic spread in networks has been studied for close to two decades [4, 3], with researchers developing many mathematical models as well as using stochastic simulation to infer properties of epidemic spread. A recent open-source Python package EoN (Epidemics on Networks) provides tools implementing many of these. These include

- Numerical solvers for many of the standard analytic models (pair-approximation, effective degree, & edge-based compartmental models).
- Tools to simulate SIS and SIR disease spread in networks. These include Gillespie simulations with efficiency improvements similar to those in [2]. However, it also provides event-driven methods which are more efficient than the Gillespie algorithm and allow for non-Markovian processes.
- Tools for visualizing epidemics, including automated generation of animations.

Documentation with installation instructions is available at http://epidemicsonnetworks.readthedocs.io/en/latest/

Simulation

Gillespie simulation is probably the most widely-method used for simulating epidemic spread on networks, but it has a number of limitations.

For a well-mixed population, the Gillespie algorithm uses the total number susceptible and infected to predict the total rate of transmissions and recovery. From this it selects the time of the next event and what that event will be. If it is a transmission, then a random susceptible individual can be chosen to become infected. If it is a recovery, then a random infected individual can be chosen to recover. This is because all susceptible or infected individuals are at the same risk of infection or recovery.

For a network, each susceptible node has its own risk of infection based on how many infected partners it has. Selecting which of the susceptible nodes becomes infected and then updating the risk level of its partners becomes challenging and introduces significant speed reductions. Although methods exist to speed this up (see [2]) these do not completely solve the problem.

An alternative method comes from noting that when a node becomes infected, we can immediately choose its time of recovery from the appropriate distribution and we can also choose the time(s) at which it transmits to its partners during its infectious period. Putting these into a priority queue and then selecting the next event from the queue results in an efficient event-driven algorithm. An advantage of this approach is that it can incorporate non-Markovian distributions which require significant effort to implement in the Gillespie algorithm [1].

Example Use

The figures show two example sessions with EoN.

References

[1] M. Boguñá, L. F. Lafuerza, R. Toral, and M. Ángeles Serrano. Simulating non-markovian stochastic processes. *Physical Review E*, 90:042108, 2014.

[2] W. Cota and S. C. Ferreira. Optimized Gillespie algorithms for the simulation of Markovian epidemic processes on large and heterogeneous networks. *Computer Physics Communications*, 219:303–312, 2017.

[3] I. Z. Kiss, J. C. Miller, and P. L. Simon. *Mathematics of epidemics on networks: from exact to approximate models*. IAM. Springer, 2017.

[4] M. E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):016128, 2002.

```
>>> import networkx as nx
>>> import matplotlib.pyplot as plt
>>> import EoN

>>> N=10**5
>>> G=nx.barabasi_albert_graph(N, 5) #create a barabasi-albert graph
>>> tau = 0.1           #transmission rate
>>> gamma = 1.0     #recovery rate
>>> #simulate SIR with 0.5% initially infected.
>>> t, S, I, R = EoN.fast_SIR(G, tau, gamma, rho=0.005, tmax = 20)
>>> plt.plot(t, I, color = 'r', label='Simulation')

>>> #The EBCM model does not account for degree correlations or clustering
>>> t, S, I, R = EoN.EBCM_from_graph(G, tau, gamma, rho=0.005, tmax = 20)

>>> plt.plot(t, I, '--', label = 'EBCM approximation', linewidth = 5)
>>> plt.xlabel('$t$')
>>> plt.ylabel('Number infected')
>>> plt.legend()
>>> plt.show()
```

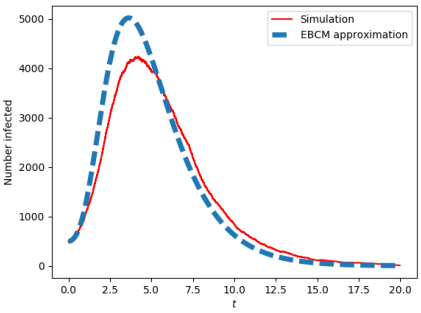


Figure 1: A comparison of simulation and a numerical approximation from an analytic model

```
>>> import networkx as nx
>>> import EoN
>>> import matplotlib.pyplot as plt
>>> G = nx.grid_2d_graph(100,100) #each node is (u,v) where 0<=u,v<=99
>>> #we'll initially infect those near the middle
>>> initial_infections = [(u,v) for (u,v) in G if 45<u<55 and 45<v<55]
>>> sim = EoN.fast_SIS(G, 1.0, 1.0, initial_infecteds = initial_infections, return_full_data=True, tmax = 10)
>>> pos = {node:node for node in G}
>>> sim.set_pos(pos)
>>> sim.display(6, node_size = 4) #display time 6
>>> plt.show()
```

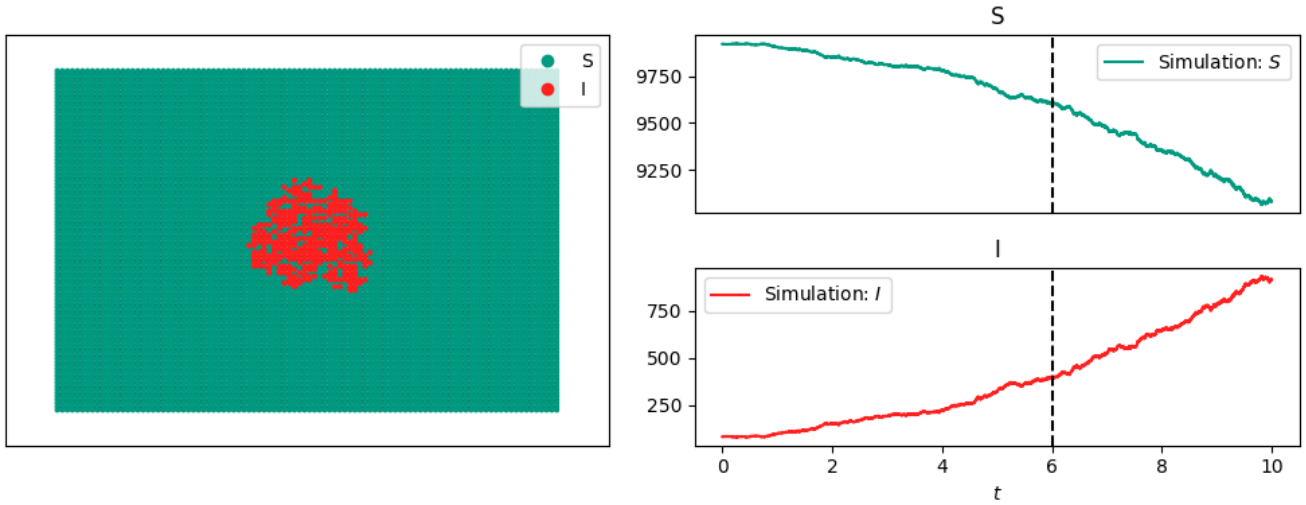


Figure 2: A demonstration of some of the visualization tools in EoN

COMPUTING THE STATISTICAL SIGNIFICANCE OF OPTIMIZED COMMUNITIES

John Palowitch

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Community detection has become a popular practice among data scientists and machine learning engineers [1]. Commonly, the goal of a practitioner applying community detection to a networked population wishes to find groups having high internal interaction and low external interaction. These groups can then be profiled and used profitably; applications have included everything from characterizing genetic regulation to building recommender systems in social networks.

Many community detection methods are based on the optimization of a network partition score, the most common of which is Newman and Girvan's "modularity" [3]. Tools to maximize modularity and other similar scores are plentiful and easy to use. However, methods for assessing the quality of the resulting partition, and its constituent communities, are less-developed. While modularity-like scores can help *compare* two partitions or two communities, independent measures of quality for discovered communities are not well-studied. In particular, it is known that modularity optimization will find strong communities in network data generated from a model with no inherent distributional community structure [5]. There is not an easy way to tell if a community returned by modularity maximization has a quality consistent with one that could have arisen by chance under a community-free model.

There are some likelihood-based approaches for determining the statistical significance of partitions estimated from community-laden models [4]. However, these approaches do not yield the significance of individual communities. This problem involves the nuance that communities from a partition returned by modularity optimization are from the tail of the distribution of randomly sampled node sets. The authors of [2] were, to our knowledge, the first to address this issue. They start with the fact that under the configuration model, an external node u's edge count to a given community C follows the hypergeometric distribution

$$d(u, C) \sim \text{hypergeo}(d(C),\ 2m,\ d(u)),$$

where m is the edge count of the network, $d(C)$ is the total degree of C, and $d(u)$ is u's degree. This distribution is used to test the "border" nodes of a community, from which a significance score is derived. An adjustment to significance scores accounting for modularity maximization is proposed.

In our work, we advance this approach in the following ways. First, we show that the adjustment for modularity maximization is actually *not* needed due to empirical distributional properties of an optimized community's out-degrees. In light of this, we develop a simplified algorithm to compute the statistical significance of individual communities that are optimized with respect to modularity.

Next, we introduce a procedure involving simultaneous testing of multiple border nodes of a community. Compared with the previous approach, our procedure increases power (the proportion of ground-truth communities declared significant) while controlling Type I error (the proportion of optimized communities declared significant under a community-free model).

Finally, we build a software package providing an accessible implementation of our method, and show that our method computes quickly on very large networks, in contrast to previous work. We re-analyze many network data sets from the network science literature to demonstrate increased ability to find significant communities in applications. We discuss how our approach can be immediately deployed in both industrial and academic pursuits.

References

[1] S. Fortunato and D. Hric. Community detection in networks: A user guide. *Physics Reports*, 659:1–44, 2016.

[2] A. Lancichinetti, F. Radicchi, and J. J. Ramasco. Statistical significance of communities in networks. *Physical Review E*, 81(4):046110, 2010.

[3] M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[4] T. P. Peixoto. Model selection and hypothesis testing for large-scale network models with overlapping groups. *Physical Review X*, 5(1):011033, 2015.

[5] P. Zhang and C. Moore. Scalable detection of statistically significant communities and hierarchies, using message passing for modularity. *Proceedings of the National Academy of Sciences*, 111(51):18144–18149, 2014.

EXPLOITING INTRA-TYPE INFORMATION IN BIPARTITE COMMUNITY DETECTION

Paola Pesántez-Cabrera, Ananth Kalyanaraman, Mahantesh Halappanavar

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary: Classical bipartite community methods only take into account inter-type edge information—i.e., edges between vertices of two different types. We present a new form of bipartite modularity (as an objective function for community detection) that can enable methods to incorporate both intra-type and inter-type edge information. Preliminary results evaluating this new form are presented.

Introduction

Bipartite graphs serve as an effective way to represent the interplay between two different data types—e.g., gene vs. disease, plant vs. pollinator, etc. (e.g., [3, 8]). Here, vertices represent the individual entities of a data type, and edges represent the interaction between the entities of two different data types. The problem of community detection, when applied to such bipartite networks, is one of co-clustering the entities of the two different types based on their inter-type interactions.

However, in many applications, we may also have *intra-type* information, which may be critical in determining the co-clustering structures [6]. For instance, considering the sequence-based similarity between genes (intra-type) could either provide the additional basis for clustering a group of genes with a group of diseases or help reveal hidden links between disease groups.

Current methods for bipartite community detection are ill-equipped to handle such intra-type information when made available. More specifically, the *modularity* metrics that they use, to measure the goodness of clustering, use only inter-type edge information. Note that a naïve way to handle both inter- and intra-type information is to simply treat the graph as a general graph and run methods that are designed for general graphs. However, intra-type may or may not carry the same weight as inter-type; furthermore, the connectivity characteristics (e.g., sparsity of edges, degree distribution) could differ between inter- and intra-type edges.

Contribution: In this paper, we present a definition of bipartite modularity that would enable bipartite community detection methods to compute clustering structures taking into account *both* inter- and intra-type edges. Our definition extends the `Murata+` definition of [7].

Notation and Definitions: Let $G = (V_1 \cup V_2, E)$ denote a bipartite graph, where V_1 and V_2 represent vertices of two different types, and an edge $e_{ij} \in E$ represents a pairwise relationship between $i \in V_1$ and $j \in V_2$. M denotes the sum of the weights of all edges in E. We define an *augmented bipartite graph* as a bipartite graph which also allows edges between vertices of the same type—i.e., $G(V_1 \cup V_2, E \cup E')$, where every edge $e_{ij} \in E'$ is such that either $i, j \in V_1$ or $i, j \in V_2$.

The goal of bipartite community detection is to partition V_1 and V_2 into a set of communities such that the members of a community are highly "related" to one another than to the rest of the network. The degree of relatedness is typically captured in the modularity of clustering.

Classical Definitions of Bipartite Modularity

Multiple bipartite modularity definitions have been proposed [1, 4, 5, 7]. However, all the above definitions focus on establishing community structures only based on inter-type information. Guimerà *et al.* [4] focuses on connectivity from the perspective of only one vertex type. Barber [1] assumes and enforces a one-to-one correspondence between the communities from the different vertex types, whereas Murata's definition [5] overcomes this limitation. During analysis, we encountered an inconsistency in Murata's definition and proposed a variant called `Murata+` defined as follows [7]:

$$Q_B = \sum_C (\mathcal{E}_{C,\psi(C)} - \mathcal{A}_C \times \mathcal{A}_{\psi(C)}) + \sum_D (\mathcal{E}_{D,\psi(D)} - \mathcal{A}_D \times \mathcal{A}_{\psi(D)}) \quad (1)$$

Here, C and D represent a community in V_1 and V_2 respectively; $\psi(C)$ denotes a D that is identified as the *co-cluster mate* of C in V_2 (similar definition for $\psi(D)$); $\mathcal{E}_{C,\psi(C)}$ represents the fraction of inter-type edges from C to $\psi(C)$ (similar for $\mathcal{E}_{D,\psi(D)}$); and A_C (or A_D) denotes the fraction of edges contributed by community C (or D).

Proposed Definition of Bipartite Modularity

Given an augmented bipartite graph $G(V_1 \cup V_2, E \cup E')$, we assume (without loss of generality) that all edges have

normalized weights. First, we define a positive weight $\alpha \in \mathbb{R}$ for using inter-type edges (implying, $1-\alpha$ for intra-type edges). We use $s(i,j)$ to denote the"similarity" score between vertices i and j of the same type. Let us consider the bipartite network formed by genes and drugs; then, $s_g(i,j)$ is a sequence-based similarity score between two genes i and j, while $s_d(i,j)$ is a structure-based similarity score between two drugs i and j. Based on the s function, we define β and ϕ factors for community C of genes as follows (for $i \neq j$):

$$\beta(C) = \frac{\sum\limits_{i,j \in C} s_g(i,j)}{\sum\limits_{i,j \in V_1} s_g(i,j)}, \quad \phi(C) = \frac{\sum\limits_{i \in C, j \in V_1} s_g(i,j)}{\sum\limits_{i,j \in V_1} s_g(i,j)}$$

Intuitively, $\beta(C)$ represents the relative intra-cluster similarity based solely on intra-type edges, whereas $\phi(C)$ is the fraction of intra-type edges (in V_1) contributed by community C.

Subsequently, we define the augmented variant of the `Murata+` modularity definition as follows:

$$\boxed{Q_B = \sum_C (\mathcal{E}'_C - \mathcal{A}'_C) + \sum_D (\mathcal{E}'_D - \mathcal{A}'_D)} \qquad (2)$$

where:

$$\begin{aligned} \mathcal{E}'_C &= [\alpha\, \mathcal{E}_{C,\psi(C)}] + [(1-\alpha)\beta(C)] \\ \mathcal{A}'_C &= [\alpha\, \mathcal{A}_C \mathcal{A}_{\psi(C)}] + [(1-\alpha)(\phi(C)\phi(C))] \end{aligned}$$

Implementation: We implemented the proposed modularity into our `biLouvain` community detection tool [7] (https://github.com/paolapesantez/biLouvain). We use a multi-level iterative scheme where vertices determine their communities at each step. We implemented two variants of how a vertex chooses its destination community C_j from its current community C_i:
Strongly constrained (SC): C_j that maximizes the modularity gain ΔQ_B and $\beta(C_j)$, such that $\beta(C_j) \geq \lambda_{V_k}$ and $\beta(C_j) > \beta(C_i)$;
Weakly constrained (WC): C_j that maximizes ΔQ_B while $\beta(C_j) \geq \lambda_{V_k}$, where λ_{V_k} is a predetermined cutoff.

Experimental Results

Test data: We experimented with an Enzyme-Interaction binary bipartite network [2] that has 1,109 nodes (664 targets and 445 drugs) and 317,841 edges (2,926 inter-type, 220,116 targets intra-type, and 94,799 drugs intra-type). We use $\lambda_{V_1} = 0.03$ and $\lambda_{V_2} = 0.25$ obtained from the average similarity scores and based on [6].

Table 1: Evaluation on an Enzyme-Interaction data set.

Alpha	**Modularity Q_B**		**Correlation Coefficient(%)**		
α	**SC**	**WC**	α **comparison**	**SC**	**WC**
0.0	2.90E-05	2.90E-05	0.0 vs. 0.1	15.88	14.76
0.2	0.259	0.281	0.2 vs. 0.3	83.44	89.34
0.4	0.381	0.420	0.4 vs. 0.5	88.39	89.41
0.6	0.509	0.562	0.6 vs. 0.7	81.19	93.32
0.8	0.647	0.718	0.8 vs. 0.9	94.08	95.41
1.0	0.869	0.869	0.9 vs. 1.0	87.31	96.76

Table 1 shows how adding intra-type information impacts the final modularity. When $\alpha = 0.0$, Q_B is small because targets and drugs form a few large communities; contrarily to when $\alpha = 1.0$. When α is increased, Q_B also increases. The SC case provides a better run-time because being more restrictive reduces the amount of work needed. Correlation coefficient percentages show the degree of conservation in the clusters obtained across different α values. α values between 0.4 and 0.6 produce approximately consistent community outputs, implying that giving roughly equal weight to inter- and intra-type edges for this input data set is desirable. When comparing clusters for $\alpha = 0.0$ vs. $\alpha = 0.1$, the major difference is a consequence of inter-type information exclusion. Finally, the less restrictive the constraint, the better correlation between clusters.

References

[1] M. J. Barber. Modularity and community detection in bipartite networks. Physical Review E, 76(6):066102, 2007.

[2] B. Chen, Y. Ding, and D. J. Wild. Assessing Drug Target Association Using Semantic Linked Data. PLOS Computational Biology, 8(7):e1002574, 2012.

[3] M. Griffith, O. L. Griffith, A. C. Coffman, J. V. Weible, J. F. McMichael, N. C. Spies, J. Koval, et al. DGIdb: mining the druggable genome. Nature Methods, 10(12):1209–1210, 2013.

[4] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral. Module identification in bipartite and directed networks. Physical Review E, 76(3):036102, 2007.

[5] T. Murata. Detecting communities from bipartite networks based on bipartite modularities. In International Conference on Computational Science and Engineering, volume 4, pages 50–57. IEEE, 2009.

[6] G. Palma, M.-E. Vidal, and L. Raschid. Drug-Target Interaction Prediction Using Semantic Similarity and Edge Partitioning. In International Semantic Web Conference, pages 131–146. Springer, Cham, 2014.

[7] P. Pesántez-Cabrera and A. Kalyanaraman. Efficient Detection of Communities in Biological Bipartite Networks. IEEE/ACM Transactions on Computational Biology and Bioinformatics, PP(99), 2017.

[8] J. Piñero, N. Queralt-Rosinach, À. Bravo, J. Deu-Pons, A. Bauer-Mehren, M. Baron, F. Sanz, and L. I. Furlong. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. Database, 2015:bav028, 2015.

PARAMETER-INDUCED UNCERTAINTY IN NETWORK INFERENCE

Arvind Prasadan, Rajmonda S. Caceres, Vijay Gadepally

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

When a network is not observed directly, the inference process often requires several parameters to be specified. After the inference, performing a task (e.g., community or clique detection) also may require parameters to be specified. There is hence uncertainty in the task results. In this work, we develop and motivate this issue, and discuss potential implications.

Problem and Motivation

While networks and network data are plentiful, it is rare that we actually observe the network itself. Rather, we often observe some set of measurements, and then perform some sort of testing or thresholding analysis to form a network. E.g., we might measure gene expressions and perform a correlation analysis to find which genes are co-expressed [3]. Furthermore, after estimating the latent graph or network, we often perform some task on the network, like community detection, clique detection, or link prediction. In both the network estimation and the task, there are often parameter choices to make, and these choices lead to vastly different networks and vastly different results. For example, if the network is formed by a statistical hypothesis test, the choice of test and the choice of false alarm control level are parameters that may be varied. If the task is community detection, the algorithm and the guess of the number of communities are parameters.

Connection to Probabilistic Graphs

Simply put, a probabilistic graph is one with uncertain edges: each possible edge exists with some probability, as opposed to existing or not. There are several recent works on probabilistic graphs, including [5, 6, 1, 2]. In this setting, most authors begin with such a graph, and perform some network task (e.g., community detection) on the uncertain network. The 'usual' solution is to use Monte Carlo methods to sample possible graphs and then aggregate the results somehow. The problem is clear, however: this approach is extremely computationally demanding: there are $2^{\binom{n}{2}}$ possible graphs on n nodes, obtained by toggling each edge. Moreover, it is not clear how to obtain a representative graph from the overwhelming number of possibilities.

The problem we consider here is closely related, since by varying some set of parameters, we obtain a family of graphs. Specifically, we are given features on the nodes of the graph, and must infer whether the edges exist. The difference, however, is that the edge probabilities in our setting are functions of the parameters, and are not independent: we instead consider a much smaller sample obtained by sampling values of a few parameters. For example, we may vary the false alarm control level between 0.001 and 0.2, and discretize the interval into a few tens of values.

An Example

Consider a dataset consisting of $n-k$ p-dimensional multivariate normal random variables $\mathcal{N}(0, \sigma^2 \mathcal{I}_p)$, and k multivariate normal random variables that are more tightly concentrated: $\mathcal{N}(0, \sigma_c^2 \mathcal{I}_p)$, where $\sigma_c^2 < \sigma^2$. If we threshold the squared Euclidean distances between the samples and form a graph (each node is a sample), we would ideally (by design) place edges between the k concentrated samples and no edges elsewhere. I.e., we have a random graph with a clique. We investigate how well we detect the clique depending on how we pick the edge threshold relative to the gap between the two variances.

Assuming that we know the variances σ^2 and σ_c^2, we might use the modularity matrix method proposed in [4]. There are hence two parameters: the False Discovery Rate (FDR) control level for forming the graph, and the FDR control level for thresholding the eigenvector and detecting the clique, hereafter referred to as the clique threshold. The first FDR level, that for forming the graph, controls the formation of edges. We want there to be edges between the clique nodes, and nowhere else. The second controls the selection of the clique nodes: we want to be able to detect the clique, but do not want to select nodes that are not in the clique.

Fixing $n = 1000$, the dimension $p = 30$, the clique size $k = \lceil 11\sqrt{n} \rceil = 348$, $\sigma = 1.0$, and $\sigma_c = 0.1$, we average our

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited. This material is based upon work supported by the Assistant Secretary of Defense for Research and Engineering under Air Force Contract No. FA8721-05-C-0002 and/or FA8702-15-D-0001. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Assistant Secretary of Defense for Research and Engineering.

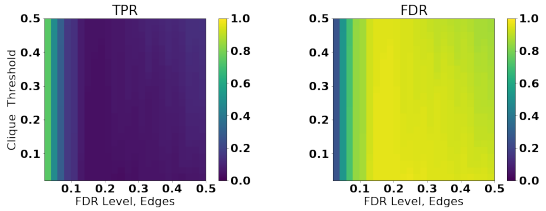


Figure 1: On the left, we plot the TPR for detecting the clique. On the right, we plot the FDR.

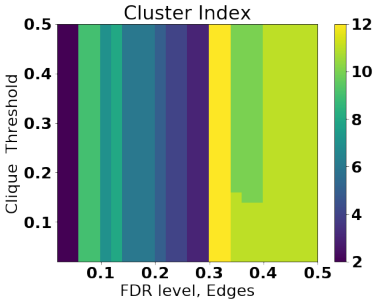


Figure 2: The cluster labels from one trial.

results over 250 trials. We choose these parameters so that the data are separable and the clique detectable. In Figure 1, we plot the True Positive Rate (TPR) for the clique selection (the proportion of the true clique nodes that are correctly selected) and the FDR for clique node selection (proportion of identified nodes that are falsely selected).

We observe that the FDR control for the edges has a very strong effect on detectability: the FDR and TPR of detection have a phase transition as this parameter varies. E.g., between an FDR level of 0.05 and 0.1, the TPR drops off sharply: a slight variation in parameter choices leads to drastic performance differences. However, the results are far less sensitive to the FDR control level for the clique detection. In a practical setting, this phase transition behavior is concerning: small parameter differences may be the difference between good results and garbage.

Additionally, we might ask how variable the results are as a function of the parameter settings. A natural way of answering this question is via clustering: by clustering the clique detection results, we hope to find that similar parameter settings are grouped together, and that studying a few examples gives a reasonably complete idea of the possible variation. Treating the clique estimates as n-dimensional binary vectors (1 if in the clique, 0 otherwise), we perform hierarchical clustering on the results. We use the silhouette method to select the number of clusters. In Figure 2, we present a heatmap of cluster membership for one realization (trial) of the data matrix. We see that for lower values of the graph-formation FDR control level, clustering is independent of the clique threshold level; for higher values, both parameters matter. Nonetheless, the clusters are cohesive and unbroken. While we present the results for only one trial, the results seen here are consistent across all of the trials we performed.

Conclusions and Future Work

We have described how an uncertain network, or, a family of networks, arises from uncertainty in parameter choices in the graph construction process. There are several real-world settings where a network is inferred, and the inference process includes several parameter choices, especially in biology. This setting has natural connections to the probabilistic graph literature, although we are faced with a far more tractable problem. We have shown a toy example of a network that is estimated, and the effect on a task that the inference process has.

We are concerned about characterizing uncertainty and propose to do it by connecting the graph construction parameters to the variation of task results. A natural procedure is the following: given a parameter of interest, we may vary it and cluster the resulting networks' task results. This procedure gives a condensed view of the range of results that the parameter choices induce. From the practical and operational perspectives, we have to worry about making this analysis feasible for large graphs. We leave this objective for future work.

References

[1] G. Kollios, M. Potamias, and E. Terzi. Clustering large probabilistic graphs. *IEEE Transactions on Knowledge and Data Engineering*, 25(2):325–336, 2013.

[2] X. Kong, P. S. Yu, X. Wang, and A. B. Ragin. Discriminative feature selection for uncertain graph classification. In *Proceedings of the 2013 SIAM International Conference on Data Mining*, pages 82–93. SIAM, 2013.

[3] D. Marbach, J. C. Costello, R. Küffner, N. M. Vega, R. J. Prill, D. M. Camacho, K. R. Allison, M. Kellis, J. J. Collins, G. Stolovitzky, et al. Wisdom of crowds for robust gene network inference. *Nature methods*, 9(8):796–804, 2012.

[4] R. R. Nadakuditi. On hard limits of eigen-analysis based planted clique detection. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 129–132. IEEE, 2012.

[5] P. Parchas, F. Gullo, D. Papadias, and F. Bonchi. The pursuit of a good possible world: extracting representative instances of uncertain graphs. In *Proceedings of the 2014 ACM SIGMOD international conference on management of data*, pages 967–978. ACM, 2014.

[6] M. Potamias, F. Bonchi, A. Gionis, and G. Kollios. K-nearest neighbors in uncertain graphs. *Proceedings of the VLDB Endowment*, 3(1-2):997–1008, 2010.

SYNTHETIC TEMPORAL GRAPH GENERATION

Sumit Purohit[1], *Lawrence B. Holder*[2], *George Chin*[1]

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Abstract

Generating a synthetic graph that is similar to a given real-world graph is a critical requirement for privacy preservation and benchmarking purposes. Various generative models attempt to generate static graphs similar to real-world graphs. However, generation of temporal graphs is still an open research area. We present a temporal-motif based approach to generate synthetic temporal graph datasets including the core algorithm, and results from two real-world use cases.

Introduction

Graphs are a natural and flexible representation of a set of entities and the relationships among them. A static graph represents a set of objects and a set of pairwise relations between them. A temporal graph is a generalization of a static graph which changes with time. Time can also be modeled as a vertex or edge label, which makes temporal graphs a special case of attributed graphs. Incorporating time into the static graphs has given rise to a new set of challenging and important problems that can not be modeled as a static-graph problem [6]. Many domains such as social networks, communication, transportation, sensor networks, co-authorship networks, and procurements can be naturally modeled as temporal graphs.

Many graph generative models are studied and developed to generate synthetic graphs. Random Model [4] and Preferential Attachment Model [2] are classic graph generative models. The Chung-Lu model provides a random model to generate power law graphs [1] using an input degree distribution. Recently Leskovec and Foloutsos [5]presented the Kronecker model based on Kronecker matrix multiplication to generate syntactic graphs that replicate multiple graph properties. All such models attempt to satisfy some global graph properties, but do not guarantee the preservation of localized structural properties.

[1]Pacific Northwest National Laboratory, Richland, WA, 99352
[2]Washington State University, Pulllman, WA, 99164, USA

This research presents a graph generative model that preserves local temporal structures while generating synthetic graphs. It defines some easy to compute temporal atomic motifs which are used to define any real-world graph. The core hypothesis of this research is that preserving local temporal-motifs is sufficient to generate synthetic graphs that also exhibit similar global graph properties of the corresponding real-world graph.

Structural Temporal Modeling

We define Structure Temporal Modeling (STM) as a process of identifying temporal-motifs in the real-world graph. We define some easy to compute atomic-motifs such those shown in Figure 1 which can characterize any given real-world graph. We guarantee that the motifs are found in mutually exclusive fashion and we do not find overlapping motifs. We define *vertex-birth-time* of a vertex as the earliest arrival time of temporal edges associated with this vertex. We define *motif-birth-time* as the earliest time at which any edge of that motif has arrived. Using these two definitions we compute the information content of a motif as the number of new and old vertices associated with the motif. This leads to multiple *temporal-atomic-motifs* for a given *atomic-motif*. For example, in Figure 1 a triangle atomic-motif is expanded to 4 temporal-atomic-motifs where 0,1,2, or 3 vertices are new (or re-used). The six atomic-motifs in Figure 1 can generate up to 20 temporal-atomic-motifs.

For each temporal-atomic-motif we also compute its *formation-time* which is the total time taken by the motif to fully form. At the same time, we also compute *average-arrival-delay* in generating each edge of the motif.

Distribution of such temporal-atomic-motifs is computed for a given real-world graph. Motif *arrival-rates* are computed by normalizing the distribution over the entire duration of the input graph. This normalized distribution is used to generate its synthetic version and the same distribution is also computed for the synthetic graph. Variation in these two distributions is used as a metric to compare quality of the synthetic graph.

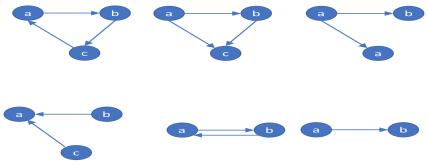


Figure 1: Atomic Temporal Motifs

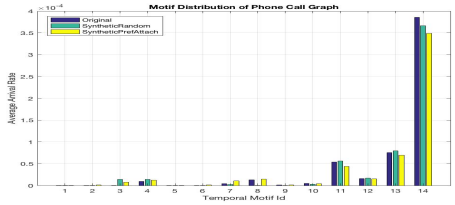


Figure 2: Synthetic Communication Network

The generator component of the STM uses the distribution to iteratively generate all the temporal motifs using arrival rates as *generation probabilities*. STM uses the information content of the motifs to decide whether to create new nodes or reuse existing nodes in the graph at a given point of time. STM also uses *formation-time* and *average-arrival-rate* to delay the formation of the temporal-motif.

Experiments

We have developed a scalable framework using Apache Spark [7] and GraphFrames [3] to compute the distribution of temporal-atomic-motifs. We have also developed a graph generator using Python (https://github.com/lbholder/graph-stream-generator) that takes the distribution as an input and generates a synthetic graph. We present results from two domains: social networks and communication networks. We were able to model one million edge graphs successfully.

Figure 2 shows the temporal motif distribution of real and synthetic snapshots of the PNNL internal communication network where each edge represents a phone communication between two persons. Similarly, Figure 3 shows the temporal motif distribution of real and synthetic Twitter graphs generated using the public API, where each edge represents a Twitter mention by source to destination. We experimented with two variations of the synthetic graph generation, random node selection and preferential node selection where a reused node is selected based on its degree. As shown in Figures 3 and 2, STM generates synthetic graphs similar to corresponding real-world graphs. It is also quantitatively evident from the very low absolute mean difference value of the motif probabilities as shown in Table below.

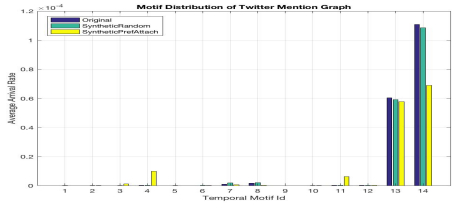


Figure 3: Synthetic Social Network

	Random	Degree
Social Network	3.7398e-07	4.5396e-06
Communication	4.4522e-06	5.3076e-06

Future Work

Future work will model multi-type graphs that increase the number of candidate temporal motifs. We will address this challenge.

Acknowledgment

We thank the DARPA Modeling Adversarial Activity program for funding this project under contracts HR0011728117, HR001178235, and HR0011729374. The associated PNNL project number is 69986. This work is also supported by the National Science Foundation under Grant No. 1646640.

References

[1] W. Aiello, F. Chung, and L. Lu. A random graph model for massive graphs. In *Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 171–180. Acm, 2000.

[2] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[3] A. Dave, A. Jindal, L. E. Li, R. Xin, J. Gonzalez, and M. Zaharia. Graphframes: an integrated api for mixing graph and relational queries. In *Proceedings of the Fourth International Workshop on Graph Data Management Experiences and Systems*, page 2. ACM, 2016.

[4] P. Erdos. On random graphs. *Publicationes mathematicae*, 6:290–297, 1959.

[5] J. Leskovec, D. Chakrabarti, J. Kleinberg, and C. Faloutsos. Realistic, mathematically tractable graph generation and evolution, using kronecker multiplication. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 133–145. Springer, 2005.

[6] O. Michail. An introduction to temporal graphs: An algorithmic perspective. *Internet Mathematics*, 12(4):239–280, 2016.

[7] A. Spark. Apache spark: Lightning-fast cluster computing. *URL http://spark. apache. org*, 2016.

Abstract: Data Fusion Reconstruction of Spatially Embedded Complex Networks

Fernando J Quevedo,[1] Erik Bollt,[1] and Jie Sun[1]
1. Clarkson University University, Potsdam, NY, USA

Inferring networks from data (e.g., estimation of brain connectivity via fMRI (7)) is important in many practical applications, especially when direct invasive measurements are infeasible. Given time series data collected on the nodes of a spatial network, the problem is to infer the underlying interaction structure of the network (6). A main challenge is that the amount of data in practice is typically small comparing to the size of the network, rendering reliable inference a difficult and sometimes impossible task. Recognizing that many real world networks are spatially embedded (16), this project utilizes such information to develop a kernel-based spatial network inference framework that significantly improves inference outcome. Our new approach enables efficient and accurate reconstruction of large spatial networks from limited data even when the exact spatial distribution of the embedded edges are not known. The results have potential impacts on biological and engineering applications where big data is being continuously collected.

Background and existing approaches. Network inference can be interpreted as an inverse problem. For large networks, the problem is ill-posed due to the small amount of data contrasting the large number of unknowns (6; 9). Since real-world networks are typically sparse, a common approach is to include a penalty/regularization term to select sparse solutions (9; 10; 11; 12). Typical choices of the parameter p in the regularization term are $p = 1$ (known as Lasso (9)) and $p = 2$ (ridge regression), or combination of both (Elastic Net (13)). Interestingly, $p = 1$ is the only value for which the solutions are parsimonious and the problem remains convex (and thus relatively easy to solve). The parsimonious property holds for $p \leq 1$ whereas the optimization problem is convex when $p \geq 1$. For sparse networks, Lasso ($p = 1$) is particularly useful for inferring sparse networks comparing to other methods that do not take advantage of sparsity.

Kernel Lasso Reconstruction approach. To incorporate the spatial embedding information, we extend the classical LASSO framework by adding a kernel function. We proposed a kernel-based Lasso (*klasso*), where the reconstruction of links that influence the state of a given i is accomplished by solving the following inverse problem as an optimization problem:

$$\theta_{klasso}^{(i,\lambda,k)} = argmin_{(\theta|\theta_i=0)}(\|X_i - X\theta\|_2^2 + \lambda \sum_{j=1}^{n} k(\|S_i - S_j\|)|\theta_j|) \tag{1}$$

Here the Tn matrix X represents time series of the n nodes (X_i is the i^th column); S_i is node is spatial location, $\|S_i - S_j\|$ is the spatial distance between i and j; $k(\cdot)$ is a kernel function; and λ is the regularization parameter. The kernel is key to impose spatial regularity. We consider a general class of power-law kernels,

$$k(\|S_i - S_j\|) = \|S_i - S_j\|^{\gamma} \tag{2}$$

with a single parameter $\gamma : \gamma = 0$ recovers the Lasso solution, while $\gamma \to \infty$ selects only connections with shortest spatial distance. Values of γ in between are balance the preference of short- versus long-distance connections.

Numerical results. We validated klasso using synthetic as well as real-world network data (1). Here we show results of *klasso* in the reconstruction of the European highway network, where

the algorithm is given time series of the number of individuals per site/node, with these individuals walk randomly on the network over time. The results shown in the figure demonstrate the excellent performance of *klasso* in accurate reconstruction of the network, especially comparing with Lasso. This significant improvement is due to klassos unique capability of fusing data (here both time series and spatial embedding information).

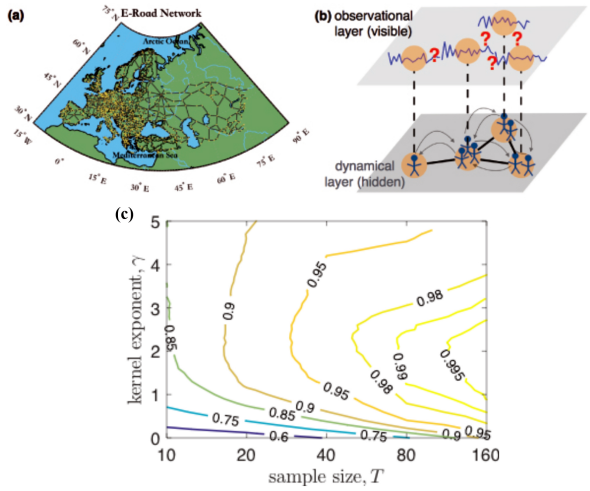


Figure 1: (a) European highway network: nodes (light yellow dots) are cities, edges (dark gray lines) are highway segments (water crossings are excluded) (1). The network has $n = 955$ spatially embedded nodes and $m = 1255$ edges. (b) Two-layer representation of the hidden dynamics on hidden network. Both the network and dynamics of individuals on the network are hidden (hidden dynamical layer), only aggregated population dynamics is measured (observational layer) (1). (c) Data-fusion reconstruction of the network by *kLasso*, shown as a contour plot of AUC values in the (T, γ) plane. Lasso solution corresponds to $\gamma = 0$, which always gives lower AUC value (worse reconstruction) than *kLasso* $(\gamma > 0)$. Here the optimal choice of γ lies somewhere between 2 and 3 (1).

I would like to explore a few important and relevant problems in the future as follow; The Non-Euclidean geometry. Our current *klasso* for spatial networks uses Euclidean metric for node-to-node spatial distance. In many applications (e.g., brain, heart, etc.) the dynamics occur on nonlinear manifolds, thus the norm in Eq(1) needs to be revisited. A good starting point is to construct appropriate geodesic through nonlinear embedding into a Euclidean space (14; 15). Furthermore, the modular and hierarchical network structure. For networks that are not spatially embedded, but have other structural characteristics such as modular or hierarchical structures, often there is information (in addition to dynamics) about node attributes. An example is occupation of people in social network. It will be useful to develop appropriate kernel functions in place of the spatial kernel as in Eq(1), to take advantage of these meta information in network reconstruction. This work brings promising to many applications i.e. in epidemic control and management, it is often challenging to find the (hidden) large-scale transmission pathways. *kLasso* can be useful by incorporating infection data together with spatial location to better track the epidemic spread pattern and source. Another application is the inference of dynamic brain connectivity (in different contexts such as resting state, active state, stress state, etc.) via fMRI data: neurons are clearly spatially embedded.

References

[1] J. Sun, F. J. Quevedo, and E. Bollt, Data Fusion Reconstruction of Spatially Embedded Complex Networks, arXiv:1707.00731.

[2] F. Altarelli, A. Braunstein, L. Dall'Asta, A. Lage-Castellanos, and R. Zecchina, Phys. Rev. Lett. **112**, 118701 (2014).

[3] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, Nat. Commun. **6**, 8502 (2015).

[4] A. S. Ambegedara, J. Sun, K. Janoyan, and E. Bollt, Chaos **26**, 116312 (2016).

[5] W. M. Lord, J. Sun, N. T. Ouellette, and E. M. Bollt, IEEE Trans. Mol. Biol. Multi-Scale Commun. **2**, 107 (2017).

[6] J. Sun, D. Taylor, and E. M. Bollt, SIAM Journal on Applied Dynamical Systems 14(1), 73106 (2015).

[7] E. Bullmore and O. Sporns, Nature Revi. Neurosci. 10, 186198 (2009).

[8] M. Rubinov and O. Sporns, NeuroImage 52, 10591069 (2010).

[9] R. Tibshirani, Journal of the Royal Statistical Society: Series B 58(1), 267288 (1996).

[10] N. Meinshausen and P. Buhlmann, The Annals of Statistics 34(3), 14361462 (2006).

[11] J. Friedman, T. Hastie, and R. Tibshirani, Biostatistics 9(3), 432441 (2008).

[12] D. M. Witten, J. H. Friedman, and N. Simon, Journal of Computational and Graphical Statistics 20, 892 (2011).

[13] Zou et. al., Journal of the Royal Statistical Society, Series B: 301320 (2005).

[14] Joshua B. et. al., Science 290, 23192323 (2000).

[15] Coifman, R.R.; S. Lafon."Diffusion maps". Applied and Computational Harmonic Analysis. 21: 530(2006).

[16] M. Barthelemy, "Spatial networks," *Physics Reports* **499**, 1–101 (2011)

EXPANDING SEED SETS WITH DEGREE CORRECTED RANDOM WALKS IN THE DEGREE CORRECTED STOCHASTIC BLOCK MODEL

Stephen Ragain

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Random walk based methods for seed set expansion are popular in practice because they are tractable [2] and empirically effective [3]. A recent result by Kloumann et al. [4] showed that Personalized PageRank (PPR) can be framed as the optimal linear classifier in the space of random walk landing probabilities from a seed node for networks arising from the Stochastic Block Model (SBM). We show an analogous result for the degree-normalized PPR and the degree-corrected Stochastic Block Model (dcSBM) by considering the space of degree-normalized random walk probabilities and show that the resulting scoring method performs well on empirical networks.

Problem Setting

We consider the problem of seed set expansion for community detection in networks. Given a directed graph $G = (V, E)$ with communities $C_1, \ldots, C_\ell$ partitioning V, we are given a seed set $S \subset C_i$ for some C_i and our objective is to recover C_i. We may also have access to $|C_1|, \ldots, |C_\ell|$, the number of nodes in each community. We adopt the notation $c(v) = i$ if $v \in C_i$.

In the stochastic block model (SBM) with $\ell \times \ell$ parameter matrix ω has for each $x, y \in V$, a directed edge from x to y occurs independently with probability $p_{xy} = \omega_{c(x)c(y)}$. A widely studied special case occurs when ω_{ij} is some fixed probability p_{out} for $i \neq j$ and some fixed probability $p_{in} > p_{out}$ for $i = j$.

The degree-corrected stochastic block model (dcSBM) allows us to capture the block structure of an SBM while specifying expected degree sequences which are heterogeneous across a block. The dcSBM has an $\ell \times \ell$ parameter matrix ω and a "friendliness" parameter vector θ. The probability that an edge from x to y exists is $\theta_x \theta_y \omega_{c(x)c(y)}$.

Personalized PageRank (PPR) is a generalization of PageRank that can be interpreted as the steady state of a random walk which at each step moves to a uniform at random neighbor of the current node with probability α or with probability $1-\alpha$ teleports to some node uniformly at random chosen from a seed set $S_0 \subset V$. For non-personalized PageRank, $S_0 = V$.

We will consider the use of linear discriminant analysis (LDA) via scoring functions $f(v) = w^T r_v(S)$ which represent linear functions of some features r_v assigned to each vertex v given the seed set S. Given two point clouds, the optimal weight vector w for separating one cloud from another is given by the difference of the centroid vectors of those clouds. When $r_v(S)$ is a length-K vector consisting of the probabilities that a random walk starting uniformly in S arrives at v in k steps for $k = 1, \ldots, K$, and the network arises from an SBM, Kloumann et al. show that the optimal weight vector is given by a geometric sequence $w = (\alpha, \alpha^2, \ldots, \alpha^K)$, Thus the LDA scoring method is equivalent to Personalized Page Rank (PPR) from the seed set with teleportation parameter $1 - \alpha$. Furthermore, in the case where ω's entries consist of p_{in} and p_{out}, they derive the optimal α in terms of p_{in} and p_{out}.

Random Walks in the DCSBM

We show that a k step random walk landing probability r_v^k from seed set S to node v in the dcSBM is tightly concentrated around its expectation, which has a convenient representation as proportional to θ_v times the expected probability that we are in v's community in k steps. This expression can be written as the solution of a simple homogeneous linear system on the communities. In other words,

$$r_v(S)^k \approx \frac{\theta_v}{\sum_{u \in c(v)} \theta_u} q^k_{c(v)},$$

where $q^k_{c(v)}$ (which depends on S) is the random walk landing probability in the SBM corresponding to the dcSBM, arising from the homogenous linear system

$$q_i^k = \sum_{j=1}^{\ell} W_{ji} q_j^{k-1},$$

where $W_{ji} = \omega_{ji} \frac{|C_j|}{n} \sum_{v \in C_j} \theta_j$ and $q_j^0 = 1(S \in C_j)$ is the initial condition that a community contains the seed set.

The core of the proof uses the independence of edges to give tight enough concentration for the number of edges from each node to each community, allowing us to bound the probability of walking from each node to each community. This gives us a centroid in the space of θ-normalized random walk probabilities for each community. Noting that $q^k_{C_i}$ comes from the homogenous linear system depending only on the community sizes and ω, so the difference in the centroids, i.e. the optimal weight vector for scoring with the normalized probabilities, is simply given by the differences in the values of the linear system for $k = 1, \ldots, K$.

When there are two equal size blocks with $\omega = \begin{bmatrix} p_{in} & p_{out} \\ p_{out} & p_{in} \end{bmatrix}$, the difference of the centroids' k-th entry is actually $\left(\frac{p_{in}-p_{out}}{p_{in}+p_{out}}\right)^k$. In this case we also have that $E[deg(v)|\theta_v] \propto \theta_v$ for all nodes v, so the scores that arise are proportional to $deg(x)^{-1}\sum_{k=1}^{K} \alpha^k r_v^k$ where $\alpha = \frac{p_{in}-p_{out}}{p_{in}+p_{out}}$, which is precisely degree-normalized Personalized PageRank.

The machinery used is a straightforward extension of Kloumann et al., but our results both provide a theoretical motivation for the already heuristically popular degree-normalized Personalized PageRank and connects it with an important class of random graphs.

Algorithm Outline

Here we give the psuedo-code for `dcsbm-rank` our scoring method for seed set expansion.

1. From adjacency matrix A, estimate $\hat{\theta}$ and $\hat{\alpha}$ and choose a maximum walk length K such that the condition number $\kappa(A^K)$ is not too big.
2. For $k = 1, \ldots, K$, compute the landing probability r_v^k of each node v from source node s drawn uniformly from S in k steps.
3. For each node v compute $r_v = \left(\frac{r_v^1}{\hat{\theta}_v}, \frac{r_v^2}{\hat{\theta}_v}, \ldots, \frac{r_v^K}{\hat{\theta}_v}\right)$
4. For $w = (\hat{\alpha}, \hat{\alpha}^2, \ldots, \hat{\alpha}^K)$, output scores $f(v) = w^T r_v$.

To output an expansion of S with a budget of b nodes, choose the b highest scoring nodes not in the seed set S.

Parameter Estimation

We need to estimate both α and θ in order to compute the scores in `dcsbm-rank`. Because all scores scale with θ, we only need to estimate θ up to a constant. Assuming equal block sizes and homogeneity of θ across blocks, this can be done with $\hat{\theta}_v = deg(v)$.

Given an estimate for θ, we can estimate α from estimates of p_{in} and p_{out}. Consider the weighted adjacency matrix $\tilde{A}_{xy} = \frac{A_{xy}}{\hat{\theta}_x \hat{\theta}_y}$. Then the moments of $\tilde{A}_{xy}$ follow the SBM model on ω and we can use moment based estimators from [1] to estimate p_{in} and p_{out} and in turn, α.

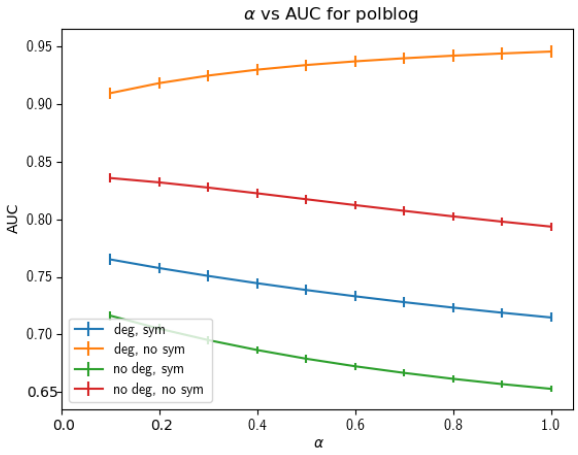


Figure 1: Area under the receiver-operator curve (ROC AUC) on the `polblogs` networks for different values of α when choosing a seed node uniform at random. 'deg' denotes the use of degree correction, and 'sym' denotes the use of symmetrization.

Applications

We demonstrate the efficacy of `dcsbm-rank` with two empirical networks, `BerkStan` and `polblogs`, by considering the AUC for classifiers expanding from a singleton chosen uniform at random from V. In figure 1 we show that the AUC under degree correction improves upon the uncorrected classifier. We use

References

[1] E. S. Allman, C. Matias, and J. A. Rhodes. Parameter identifiability in a class of random graph mixture models. *Journal of Statistical Planning and Inference*, 141(5):1719–1736, 2011.

[2] B. Bahmani, A. Chowdhury, and A. Goel. Fast incremental and personalized pagerank. *Proceedings of the VLDB Endowment*, 4(3):173–184, 2010.

[3] I. M. Kloumann and J. M. Kleinberg. Community membership identification from small seed sets. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1366–1375. ACM, 2014.

[4] I. M. Kloumann, J. Ugander, and J. Kleinberg. Block models and personalized pagerank. *Proceedings of the National Academy of Sciences*, page 201611275, 2016.

Greenberg-Hastings dynamics on a small-world network: the effect of network disorder on the collective extinct-active transition

Leonardo Reyes[1,*], Anthony Ramos[1], Fernando Zhapa[2] and Miguel Pineda[3]

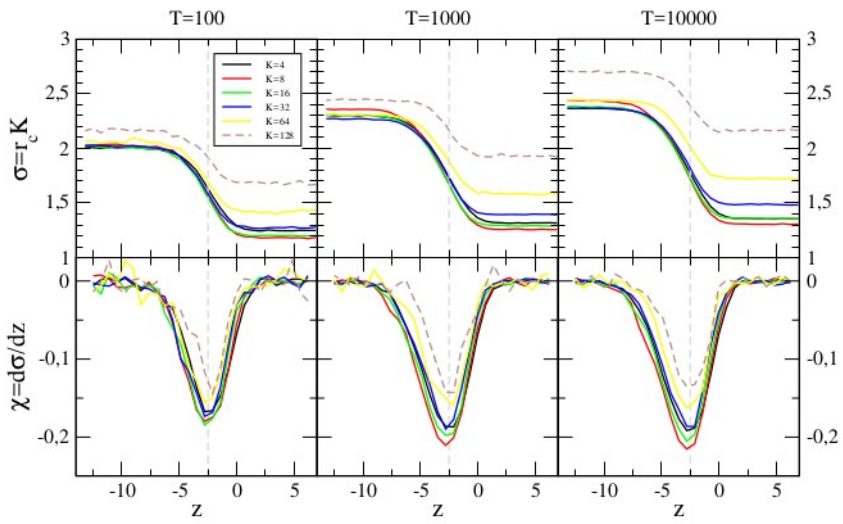
1 YachayTech University, School of Physical Sciences and Nanotechnology, Ecuador
2 YachayTech University, School of Mathematical Sciences and Information Technology, Ecuador
3 Dept of Chemical Engineering, Faculty of Engineering Science, UCL, London

* corresponding author. Email: lreyes@yachaytech.edu.ec

Abstract:
We present a numerical study of a reaction-diffusion model on a small-world network. We characterize the average activity in the system F_T after T time steps and the transition from a collective extinct state to an active state in parameter space. We found that F_T does not depends on disorder in the network if the transmission rate r or the average coordination number K are large enough. The collective extinct-active transition can be induced by changing two parameters associated to the network: K and the disorder parameter p (which controls the variance of K). We can also induce the transition by changing r, which controls the threshold size in the dynamics. Our results are relevant for complex systems that operate close to critical points. We discuss how glassy behaviour appears within our model.

Keywords: Complex Networks, Complex systems, Cellular automata, Phase transitions

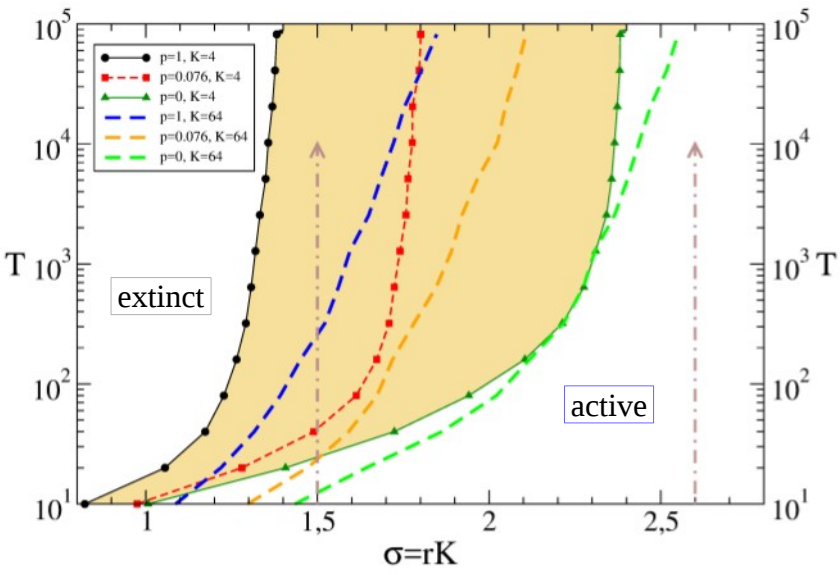
As complex systems may benefit from operating near critical points it is of interest to study under what conditions a critical state is achieved. As a prototype model we choose to study the Greenberg-Hastings (GH) cellular automata on a Watts-Strogatz (WS) Small-World network. GH is a three state discrete model for reaction-difussion that was used in the context of the Belousov reaction: it's a model for excitable dynamics. In it's simplest version this model includes only one parameter, the transmission probabilty r, which controls the threshold size in the dynamics. With the WS model we study the effect of network disorder in the collective extinct-active transition. In reference [1] it was shown that, in order to reproduce experimentaly obtained neural patterns, a variant of a GH model implemented in the human connectome must be tuned to operate near a critical point.
It can be shown that in a mean field treatment of the GHWS model it reduces to the dynamics of mean field nodes coupled to the mean activity F (which depends on the state of every node), with a coupling to the mean constant $\sigma=rK$. Here the activity F is defined as the fraction of active nodes in the system. σ can also be interpreted as the minimal number of effective neighbours needed to be in a collective active state. Since the crossover from a regular lattice to a smal-world one in the WS model depends on the product Np [3], where N is the number of nodes and p is the rewiring probability, in order to characterize the effect of disorder in the dynamics we will need, in general, to consider several orders of magnitude for p; because of this we characterize the disorder in the network with the variable $z=\ln[p/(1-p)]$, with $z\in(-\infty,+\infty)$. In the following figures we will show the behaviour of the model, with respect to the extint-active transition, as a function of σ, z and K.



Points in parameter space that are above the sigmoid curves are in a collective active state. Disorder favours activity, in the sense that reduces the minimal coupling needed for being in a active state. The transition curves depends on K and on T. For very ordered (small z) and for very disordered (large z) networks we can cross the transition curves only by changing the coupling parameter σ. For intermediate values of disorder we can cross the transition curves by changing the disorder in the network, while keeping the coupling parameter constant.

We characterize the susceptibility of the frontier to changes in disorder by introducing the quantity $\chi=d\sigma/dz$. In the figure above it is shown that there is an optimal value of z for which the transition can be made with a minimal change in disorder. We obtain $z_c\approx-2.5$ (equivalent to $p_c\approx0.08$) independently of K and T. A mean field simulation of this system shows qualitative similar results. This should be relevant in the context of adaptation, since the system can switch

between qualitatively different collective states with a minimal internal reorganization, and also in the context of communication between networks [7].



We define an ensamble of constant coupling σ, with any value of disorder. In the figure we show the evolution of two of these ensambles: σ=1.5 and σ=2.6. These ensambles are equivalent to horizontal lines in the previous figure. For σ=1.5, for example, the evolution of the ensamble hits first the curve associated to p=0 and it will not hit the curve for p=1 after very long simulation times T, if K is small. This is reminiscent of glassy behaviour: the ensamble will evolve keeping a finite activity for very long times and, in average, the ensamble will become more disordered since the more ordered members of the ensamble will go extinct first [5,7]. If K is large the whole ensamble σ=1.5 will go extinct after a short time. The evolution of the ensamble σ=2.6 will be to be in a collective active state for very long times, if K is small.

It has been reported [10,11] that when considering two interacting networks the whole system can be found in two different states. If the coupling between the networks, which we charaterize with a parameter q, is small then we can find that one network is in an active state while the other is in an extinct state. If the coupling between the networks is large, then the two networks will be in the same collective state, depending on parameters. We are currently working on the simulation of such an scenario within the results presented here. We will adjust the coupling between the two networks so that we are close to q_c. Near q_c: what is the influence of disorder in the overall behaviour?

References:

1) A. Haimovici, E. Tagliazucchi, P. Balenzuela, D. R. Chialvo, Phys. Rev. Lett. 110 (2013) 178101.
2) D. J. Watts, S. H. Strogatz, Nature 393 (1998) 440.
3) A. Barrat, M. Weigt, Eur. Phys. J. B 13 (2000) 547–560.
4) D. S. Bassett, E. Bullmore, Neuroscientist 12 (2006) 512.
5) A. Scala, L. A. Nunes Amaral and M. Barthélémy, Europhys. Lett., 55 (4), pp. 594–600 (2001).
6) J. M. Greenberg, S. P. Hastings, SIAM Journal on Applied Mathematics 34 (1978) 515–523.
7) Reyes. L., arXiv:1505.00182, http://arxiv.org/abs/1505.00182 (2015)
8) Adrian Daerr and Stephane Douady, Nature 399, 241-243 (20 May 1999);
A. Smart, P. Umbanhowar, J. Ottino, EPL 79 (2007) 24002;
L. Staron, F. Radjai, Phys. Rev. E 72 (2005) 041308.
9) John M. Beggs and Nicholas Timme, Front Physiol. Vol. 3, pp 163, 2012.
10) Mark Dickison, S. Havlin, and H. E. Stanley, Phys. Rev. E 85, 066109 – Published 8 June 2012.
11) Meng Liu *et al*, PLoS One. 2015; 10(3): e0120701.

THE NETWORK TOPOLOGY OF LOCALLY INTERACTING AGENTS AND THE DISTRIBUTION OF POSSIBLE AGENT ACTIONS IN AN ECONOMY WITH A FIXED AGGREGATE FEATURE

Janelle Schlossberger

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

This work nuances theories of aggregation in economic systems, and it develops a set of tools for closed-form analysis of complex economic systems. This work considers an economy with N networked agents and a fixed aggregate feature, and it asks whether there exists a non-degenerate distribution of possible paths along which the economy can evolve. For every feasible population size and network topology, this work maps the underlying network structure to a distribution of possible paths for the system.

Theoretical Framework

We consider economies with N networked agents and an aggregate feature. Within an economy, each agent in the population has a binary-valued attribute, and there are $n \leq N$ agents with the unit value of the attribute. Decision-making for each agent depends on the local relative frequency of that attribute's unit value; network position determines the local environment for each agent. The population sum of agents' binary-valued attributes, n, or equivalently, the *global* relative frequency of the attribute's unit value, $\frac{n}{N}$, constitutes the aggregate feature for the economic system. This work seeks to characterize the distribution of possible paths along which an economy can evolve given its aggregate feature.

The existence of a non-degenerate probability distribution of possible paths arises if the economic system exhibits configuration dependence. A configuration is an allocation of the binary-valued attribute among agents, so two configurations are distinct if they respectively have different subsets of n agents possessing the attribute's unit value. For an economy whose aggregate feature is $\frac{n}{N}$, there are combinatorially many possible configurations, ${}_NC_n$, consistent with that system's aggregate feature. Notice that as the configuration changes, the subset of nodes on the network with the attribute's unit value adjusts, so the local features of an agent's neighborhood change. The local relative frequency of the attribute adjusts, which can consequently make an individual agent's action change. The adjustment of individual agents' actions can cause the aggregate action to change. Depending on the network's topology, there can be strong dependence of the aggregate action or path of the economic system on configuration.

This work does indeed ultimately study the distribution of possible paths along which the economy can evolve for a fixed aggregate feature. However, before delving into this pursuit, this work first dedicates much space towards the characterization of a slightly abstracted, intermediate quantity: the local relative frequency of the attribute. Depending on the ultimate action of interest, the local relative frequency of the attribute gets computed differently, but its interpretation is always the same: it is the relative prevalence of the attribute's unit value. Given the network and the system's aggregate feature, there is a distribution of possible local relative frequencies of the attribute. There can potentially be significant divergence of the local relative frequency of the attribute from the attribute's global relative frequency. The scope for such divergence depends on the properties of the network. To the extent that this local relative frequency deviates in either direction from the attribute's global relative frequency, there can be significant variation in the economy's path, holding the aggregate feature of the system fixed. It is this distribution of local relative frequencies of the attribute that serves as an input for constructing the distribution of possible paths along which the economy can evolve. The distribution of local relative frequencies of the attribute must be non-degenerate in order to have a non-degenerate distribution of paths along which the economy can potentially evolve for a fixed aggregate feature.

Findings

For every feasible network topology, population size, N, and global prevalence of the binary-valued attribute, n, this work can characterize the shape and features of the distribution of possible local relative frequencies of the attribute. It determines those network topologies for which configuration is irrelevant and the distribution of possible

paths for the economy is degenerate, and it determines those network topologies for which the variance of the distribution of possible local relative frequencies of the attribute is maximal. This work shows how the topology of the underlying network can shape the higher-order statistical features, including skewness and heavy-tailedness, of system-level probability distributions. This work undertakes sensitivity analysis, and it analytically characterizes the effects of network perturbations on the distribution of possible local relative frequencies of the attribute. This work characterizes this distribution of possible local relative frequencies of the attribute when every configuration is equally likely to occur and when every configuration occurs with some arbitrary probability. This work then employs results obtained about the distribution of possible local relative frequencies of the attribute so that we can characterize the distribution of agent actions and/or paths for the economy given the system's aggregate feature.

Implications

This work nuances theories of aggregation. If decision-making for every agent depends on the global relative frequency of the attribute, we are able to define a representative agent. The decision-making behavior for this representative agent aggregates the individual decision-making function of each agent in the population. Then, holding the aggregate feature of the economy fixed, there is just one path along which the economy can evolve. For every possible configuration of the binary-valued attribute, there is no adjustment in the aggregate action because each individual agent's action depends on the fixed global prevalence of the attribute. In such a setting, we can condense the economic system with N agents into one with a single representative agent and an aggregate feature.

In general, however, it is not possible to condense the economy with a fixed aggregate feature into a system with a single agent. When agents make decisions based on their local environments, as determined by an underlying network structure, there is usually a non-degenerate distribution of paths along which the economy can evolve. Reshuffling attributes among networked agents changes the economy's path. Instead, we have a mapping from a fixed aggregate feature to a distribution of possible paths; the aggregate feature is not sufficient for determining how the system will evolve, and two economic systems with the same aggregate feature may end up evolving differently.

This work moreover contributes to the analysis of complex systems, and more specifically, complex economic systems. Agent-based interactions arise when agents use the local relative frequency of the attribute among neighbors to inform their decision-making. Depending on the setting, agent-based interaction can also arise if agents communicate their local relative frequency of the attribute to their neighbors and use an updated value to influence their decisions. The theoretical setting for the present work therefore features interaction among agents, with a network structure capturing the full set of interactions. This work develops a set of tools for the closed-form analysis of such complex systems. These tools allow us to collapse the complexities of agent-based interaction into a simple probability distribution that prescribes the way that the system will evolve. When an economy has a fixed aggregate feature and agent-based interaction, the system is quite complicated, and there are multitudinous ways that the system can evolve. Given the action of interest, this work maps the combinatorially many possible configurations consistent with the system's aggregate feature into a simple closed-form probability distribution of possible local relative frequencies of the attribute, from which we can construct the distribution of possible actions.

Application

The findings of this work and the tools developed therein can be applied to several different settings. Here, we focus on one specific application: locally formed macroeconomic sentiments and political election outcomes, particularly the outcome of the 2016 U.S. presidential election. We apply the body of theoretical results developed in this work towards understanding the level of macroeconomic sentiment in an economy and quantifying variations in macroeconomic sentiment absent changes in economic fundamentals for a very large population of agents. We show how the underlying interaction structure among agents in an economy shapes the capacity for there to exist non-fundamental swings in aggregate sentiment even at extremely large population sizes. This work therefore provides a microfoundation for animal spirits. Such sufficiently large variation in macroeconomic sentiment for fixed economic fundamentals can change the outcome of the election and thus change the path along which the economy evolves. Furthermore, when configurations are not equally likely to occur, there can be particularly strong divergence of macroeconomic sentiment away from a level that is commensurate with the economy's fundamentals.

GROWING SIGNED NETWORKS

Leah Shaw, Corynne Dech, Shadrack Antwi

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Scale-free networks grown via preferential attachment have been used to model real-world networks such as the Internet, citation networks, and social networks. Here we investigate signed scale-free networks where an edge represents a positive or negative connection. We present analytic results and simulation for growing signed network models.

Details

Growing a network with preferential attachment, where the probability for a new node to connect to an existing node is proportional to the node's current degree, leads to a scale-free degree distribution [1]. However, in a real-world network, connections might not all have the same effect. For example, some social connections might be positive, while others might represent animosity. In a signed network, each link is labeled either positive or negative. We define the positive degree to be the number of positive links a node participates in and the negative degree to be its number of negative links. Signed networks have been used, for example, to identify top users and "trolls" in an online social network [3].

Most studies of signed networks have focused on structural balance, such as looking for cycles with the relationship that "the enemy of my enemy is my friend" [4, 2, 3]. Here, we focus instead on the process of growing a signed network with a modified form of preferential attachment. We discuss several options for preferential attachment in a signed network.

In the simplest case, separate preferential attachment, new positive links attach to nodes with probability proportional to their positive degree, and negative links attach proportional to their negative degree. We analytically predict the joint distribution for a node's positive and negative degree and compare this with simulations of such networks. More complicated cases cannot be solved analytically by our method, but we use simulations to compare several other attachment rules.

Finally, we propose a method to measure preferential attachment in a real-world signed network by modifying a method developed by Pham *et al.* called PAFit maximum likelihood estimation [5]. With this proposed method, it may be possible to distinguish which of our attachment rules most closely matches reality.

References

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 10 1999.

[2] T. Hiller. Friends and enemies: A model of signed network formation. *Available at SSRN 2371249*, 2013.

[3] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: Mining a social network with negative edges. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 741–750, New York, NY, USA, 2009. ACM.

[4] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1361–1370. ACM, 2010.

[5] T. Pham, P. Sheridan, and H. Shimodaira. Pafit: A statistical method for measuring preferential attachment in temporal complex networks. *PLOS ONE*, 10(9), 9 2015.

EVIDENCE OF STEREOTYPICAL FUNCTIONAL NETWORK MOTIFS ACROSS TEMPORAL SCALES IN VERY HIGH-DIMENSIONAL BRAIN SIGNALS

Catherine Stamoulis, Department of Medicine, Harvard Medical School

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Functional brain networks vary dynamically with time, as the brain recruits local and long-distance networks to process thousands of inputs from the outside world and respond to the demands of countless behaviors and cognitive processes. Although potential combinations of network elements are theoretically infinite, there is growing evidence that across scales of organization a small number of network motifs are the building blocks of observed network topologies measured with imaging and electrophysiological data. Here we present evidence of the existence and structure of these motifs, and their relationship to network stability, in macroscale human brain signals collected over long periods of time in naturalistic (unsupervised) settings.

Emerging network motifs as a function of stability

Small-world and scale free topologies are an inherent characteristic of the organization of many real-world networks, including the brain's structural and functional circuitry. These topologies facilitate optimally efficient processing of neural information and the brain's flexibility and rapid response to cognitive demands and myriads of inputs from the external world. To date, the building blocks of these topologies (network motifs), their dynamic stereotypy and relationship to other network properties or system stability remain only partially understood. Unrelated theoretical work has shown that specific network patterns emerge as a result of a dynamic system's convergence to a stable state or configuration that favors stability [1].

Using very high-dimensional human electrophysiological signals, collected continuously over long periods of time from a relatively large number of children (n = 52) with invasive electrode arrays covering different parts of the brain, this study investigated the emergence of functional network motifs and their relationship to network stability. Contraction theoretic measures, particularly contraction loss estimated dynamically from the weighted adjacency matrix of time-varying brain networks, were used to identify sub-network patterns (motifs) that occurred repetitively over time, independently of the area of the brain being spatially sampled.

A small number of functionally active nodes (typically a set of 3-4 interconnected nodes) forming subgraphs that were estimated consistently across temporal scales and brain regions. The occurrence of these motifs was significantly associated with the stability of the dynamically estimated networks, estimated from the largest eigenvalue of the corresponding adjacency matrices. These preliminary results suggest network-level modularity and parsimony of the brain's neural circuitry, involving theoretically infinite combinations of relatively small number of motifs. Using data-estimated motifs, simulations are necessary to investigate their emergence as a function of network stability.

Meeting hashtag: #SIAMNS18

References

[1] MT Angulo, YY Liu, JJ Slotine, Network motifs emerge from interconnections that favour stability, *Nat Physics*, 11:848-852, 2015.

HOW DO GRAPHS EFFECT GLOBAL STABILITY OF SYNC IN A SIMPLIFIED KURAMOTO MODEL?

Yury Sokolov, G. Bard Ermentrout

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

The Kuramoto model on sparse graphs with identical natural frequencies and binary coupling parameter always admits a synchronized state (sync) as a stable equilibrium. However, do there exist others, i.e., when sync is not globally stable? We provide a partial answer to this question. We show that our approach works where methods like master stability function fail to distinguish graphs with different global stability properties of sync.

Introduction

We consider a simplified Kuramoto model on a sparse graph G of order n, that is, a system of n coupled oscillators, where an oscillator j is connected with a unit coupling to its neighbors in G, and the natural frequency of every oscillator is zero. Let the state of the system be $\theta = (\theta_1, \dots, \theta_n)^T$, then the dynamics is governed by differential equations of the form

$$\dot{\theta}_j = \sum_{i \sim j} \sin(\theta_i - \theta_j), j \in \{1, \dots, n\}, \tag{1}$$

where the sum is taken over all neighbors i of j in G.

The synchronized state $(0, \dots, 0)^T$ is always a stable equilibrium of (1). However, as it was observed in, e.g., [3], if G is a cycle graph with sufficiently many vertices then (1) has also twisted states - stable equilibria different from sync. This observation have raised question how does the basin of attraction of sync depend on the underlying graph of the network of coupled oscillators?

This question is about global stability of sync, and unfortunately, there have not been tools developed to address this question in full generality. Moreover, one of the methods that has been used in the community, the master stability function introduced in [2], cannot separate between graphs with the same synchronizability but different global properties of sync as we show based on examples of cospectral regular graphs. In order to overcome this, we suggest a new way to describe how the structure of underlying sparse graphs of coupled oscillators influence global stability of sync and lead to creation of phase-locked solutions of (1).

Results

In [1], we reformulate the above question of global stability of sync in Kuramoto model in terms of electrical circuits. We show that all phase-locked solutions of (1) are defined just by subgraphs, and describe an algorithm to search for them. We illustrate that Kuramoto model defined on synthetic and real nonsocial (large) networks tends to have phase-locked solutions, which reduce the basin of attraction of sync.

More precisely, we postulate that every phase-locked solution of (1) defined on a sparse graph of order n has a core that generates it. The core of a phase-locked solution consists of a subset of cycles in the graph. In order to find out whether a given graph has a phase-locked solution, we partition the vertex set into communities that form a cyclic graph clustering. If such a partition exists then we can determine the core of a phase-locked solution, and as a consequence, we obtain that sync is not globally stable. Our technique allows to find a point in the basin of attraction of a phase-locked solution in the vicinity of equilibrium. This is done by a linear transformation that is defined by the Moore-Penrose inverse of the graph Laplacian performed on an n-dimensional vector defined by the core.

References

[1] Ermentrout G.B. and Sokolov Y. When is sync globally stable in the sparse networks of identical kuramoto oscillators? *in preparation*, 2018.

[2] Barahona M. and Pecora L.M. Synchronization in small-world systems. *Phys. Rev. Lett.*, 89(5):054101, 2002.

[3] Wiley D.A. Strogatz S.H. and M. Girvan. The size of the sync basin. *Chaos*, 16:015103, 2006.

A TOY MODEL OF PREFERENTIALLY-ATTACHED NETWORKS WITH BREAK-UPS

Hiroshi Toyoizumi, Takahiro Shimomura

SIAM Workshop on Network Science 2018
July 12-13 · Portland

Summary

Break-ups in the scale-free network are studied. We study the basic features such as the dynamics of the number of break-ups, and the size of the fragments. We also discuss the extension of our toy model to $M/M/\infty$ stochastic dynamic networks with joins and leaves.

Backgrounds

Most of the current network models are focused on generating connected graphs [8], which is ideal for studying spread dynamics. However, the focus is shifting to the segregation or how closed communities arise in the network [6, 4]. We use a simple toy model of network with preferential attachment to study the segregation phenomena. In scale-free networks [1, 5, 3], nodes join the network and attach to the existing nodes preferentially with their degree, and the network evolves progressively. We use a particular type of scale-free networks [3] so that it may have break-ups.

Preferential Attachment with Possible Break-ups

Consider an evolving social network with arriving nodes. The nodes in the social network are connected with pairwise relationships called links. At the time of its arrival, a new comer selects and establishes a new link with existing nodes stochastically and more preferentially if they have more links. In addition to this preferential attachment, the new comer may select to be independent and start a new fraction in the network. The fraction founded by one individual eventually obtains followers and grows by the mechanism of the stochastic preferential attachment (see Figure 1).

Suppose one node arrives at the social network at each time instant $t = 1, 2, \ldots$. Note that we use t for both time and node index. Let $G(t) = (V(t), E(t))$ be the graph of the network. At time t, a new 1-degree node t joins to $V(t-1) = \{1, 2, \ldots t-1\}$, and one new link is added. The destination node of the new links are selected from $V(t)$ with probability proportional to its degree. Here we allow the node t to have a self-loop by assuming it has already one outgoing link. When the new link from the node t is self-loop, a break-up occurs at time t, and its probability is given by

$$P(\text{a break-up starts at time } t) = \frac{1}{2t-1}.$$

The break-ups originated at the arrival of new nodes cannot be repaired.

The network constructed by the above preferential attachment is indeed random paring network and a scale-free network [3, 5].

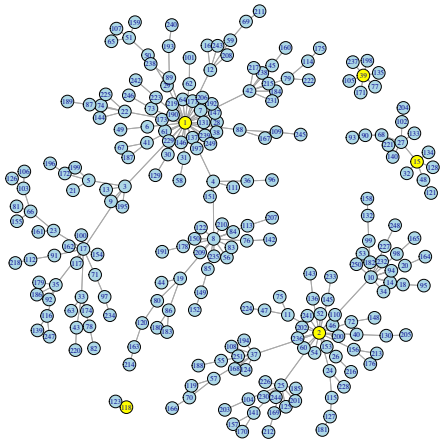


Figure 1: An example of the network $G(250)$. Yellows are the nodes who start break-ups.

The Number of Break-Ups

The number of break-ups $B(t)$ is given by a simple random walk and has the logarithmic growth.

Theorem 1. *Let $B(t)$ be the number of break-ups at time t of the network $G(t)$. Then,*

$$E[B(t)] = \sum_{k=1}^{t} \frac{1}{2k-1} \approx \frac{1}{2}\log(2t+1).$$

Break-ups can occur any time but the growth of its fraction may depend on the time it borns. It turns out that the growth of the size of the fraction is linear and the slope is inversely proportional to the break-up time (see Figure 2).

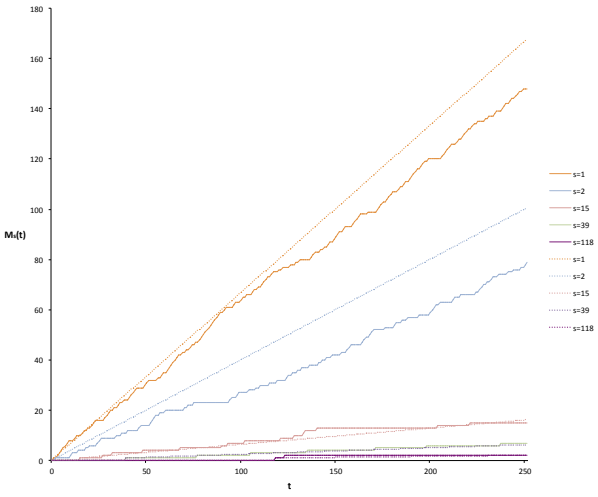


Figure 2: An example of the fraction size evolution $M_s(t)$ with the estimates (dotted lines) obtained in Theorem 2. The resulted fragmented network at $t = 250$ is shown in Figure 1.

Theorem 2. *Let $M_s(t)$ be the size of the fraction started by the node s at its arrival of $G(s)$. Then,*

$$E[M_s(t)|\textit{a break-up occurs at time } s] = \frac{2t+1}{2s+1}.$$

Break-ups in an $M/M/\infty$ Stochastic Dynamic Network

Consider a dynamic network on continuous time with leaves as well as joins. Let $C_1, C_2, \ldots$ be the nodes in the arriving order. Starting from the empty network $H(0)$, let $H(t) = \{V(t), E(t)\}$ be the network at the time t. The nodes in $V(t)$ are determined as the dynamics of $M/M/\infty$ queue, i.e., nodes arrive at the network according to a Poisson process with the rate λ and each node stays in the network for an independent exponential time with the mean $1/\mu$. Thus, $N(t)$, the number of nodes at time t, is Poisson random variable with the mean $\rho = \lambda/\mu$ in the steady state.

Given $H(t-)$ at the time of its arrival, the arriving node selects a node in $V(t)$ to establish the link between them according to the preferential attachment. At the time of leave, the leaving node C withdraws its outgoing link from the links $E(t-)$ and the incoming links to C are relocated to the next node in the arriving order existed in $V(t-)$. Note that C is the last in $V(t)$, then there is no incoming link to C.

We can show that the network $H(t)$ can be given by $G(t)$ of the size t with neglecting the leave of nodes. By

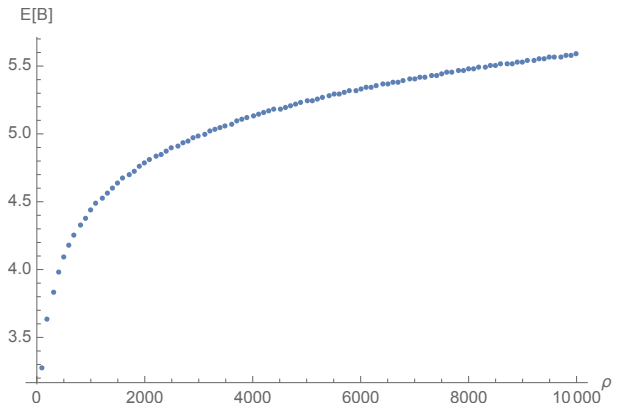


Figure 3: The expected number of break-ups $E[B(t)]$ in the $M/M/\infty$ dynamic stochastic network $H(t)$.

combining this result with Theorem 1, we obtain the following estimation of the break-ups in our toy model as depicted in Figure 3.

Theorem 3. *The number of breakups $B(t)$ in the $M/M/\infty$ stochastic network $H(t)$ in the steady state with $\rho = \lambda/\mu$ can be given by*

$$E[B(t)] = e^{-\rho} \sum_{n=1}^{\infty} \sum_{k=1}^{n} \frac{1}{2k-1} \frac{\rho^n}{n!}.$$

References

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 10 1999.

[2] N. Berger, C. Borgs, J. Chayes, and A. Saberi. On the spread of viruses on the internet. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 301–310. Society for Industrial and Applied Mathematics, 2005.

[3] B. Bollobás and O. Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1):5–34, 2004.

[4] P. S. Chodrow. Structure and information in spatial segregation. *Proceedings of the National Academy of Sciences*, 114(44):11591–11596, 10 2017.

[5] R. Durrett. *Random graph dynamics*, volume 20. Cambridge Univ Pr, 2007.

[6] X. Han, S. Cao, Z. Shen, B. Zhang, W.-X. Wang, R. Cressman, and H. E. Stanley. Emergence of communities and diversity in social networks. *Proceedings of the National Academy of Sciences*, 114(11):2887–2891, 2017.

[7] M. Molloy and B. Reed. The size of the giant component of a random graph with a given degree sequence. *Combin. Probab. Comput*, 7:295–306, 1998.

[8] X. F. Wang and G. Chen. Complex networks: small-world, scale-free and beyond. *IEEE circuits and systems magazine*, 3(1):6–20, 2003.