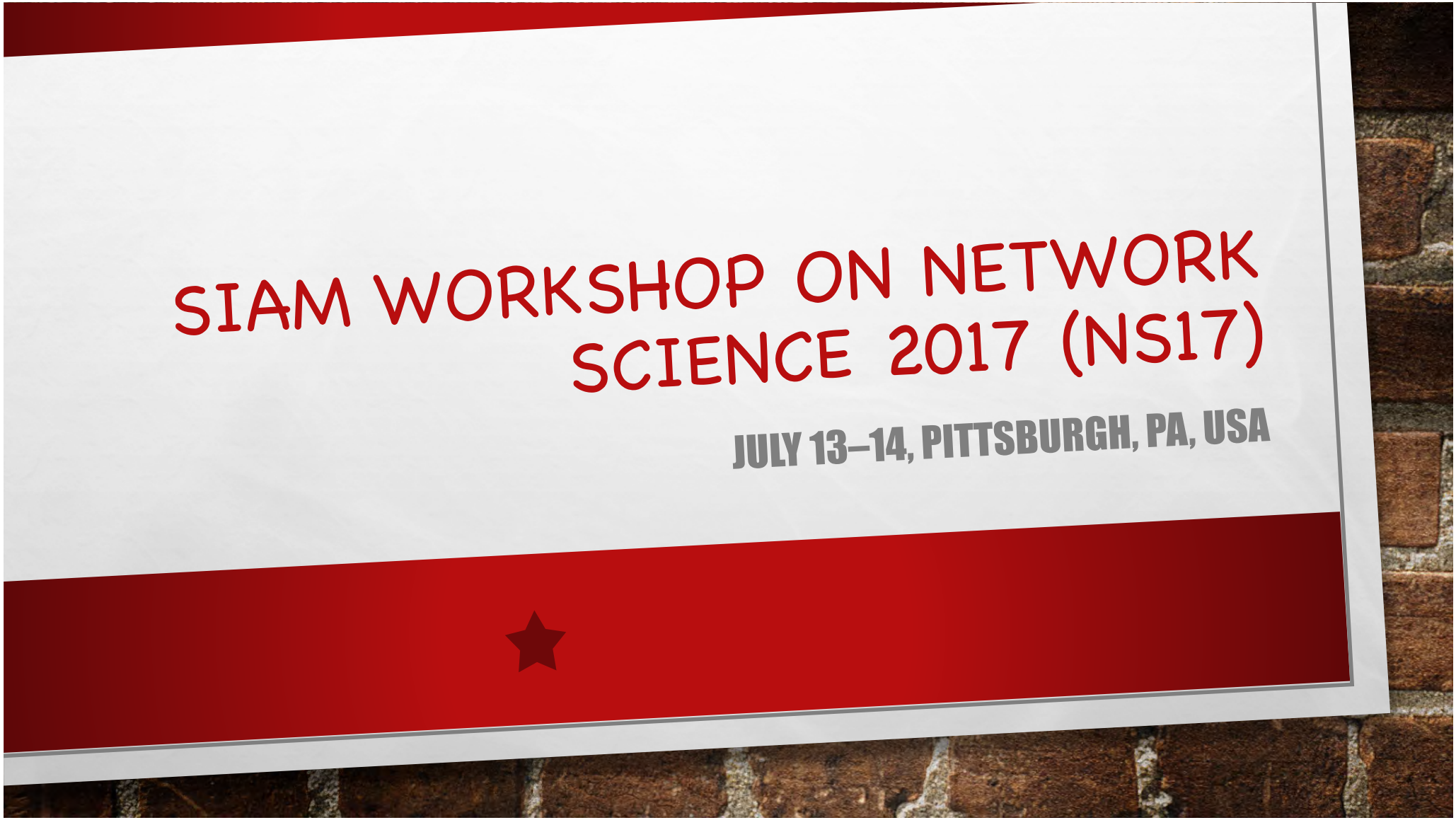

SIAM WORKSHOP ON NETWORK SCIENCE 2017 (NS17)
JULY 13–14, PITTSBURGH, PA, USA

TALK ABSTRACTS

(in order of presentation)

INVITED TALK 1: MARK NEWMAN, UNIVERSITY OF MICHIGAN

Estimating structure in networks from complex or uncertain data

Most empirical studies of networks assume simple and reliable data: a straightforward edge list or adjacency matrix that reflects the true structure of the network accurately and completely. In real life, however, things are rarely this simple. Instead, our data are complex, consisting not only of nodes and edges, but also weights, values, annotations, and metadata. And they are error prone, with both false positives and false negatives, conflated nodes, missing data, and other sources of uncertainty. This talk will discuss methods for estimating the true structure and properties of networks, even in the face of significant uncertainty, and demonstrate some applications, particularly to social and biological networks.

HOW TO EFFICIENTLY REVEAL COMMUNITY STRUCTURE IN MEMORY AND MULTILAYER NETWORKS

Christian Persson, Ludvig Bohlin, Daniel Edler, and Martin Rosvall,*

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Key message

Comprehending complex systems often require modeling and mapping of higher-order network flow representations to simplify and highlight important dynamical patterns. However, diverse complex systems require many different higher-order network representations, including memory and multilayer networks, to capture actual flow pathways in the system. Here we show that various higher-order network flow representations, including memory and multilayer networks, can be represented with sparse memory networks to efficiently reveal community structure in higher order network flows.

Abstract

To comprehend the flows of ideas or information through social and biological systems, researchers develop maps that reveal important patterns in network flows. In practice, network flow models have implied conventional first-order dynamics, but recently researchers have introduced higher-order network flow models, including memory and multilayer networks, to capture patterns in multi-step pathways. Higher-order models are particularly important for effectively revealing actual, overlapping community structure, but higher-order flow models suffer from the curse of dimensionality: their vast parameter spaces require exponentially increasing data to avoid overfitting and therefore make mapping inefficient already for moderate-sized systems.

To overcome this problem, we introduce an efficient cross-validated mapping approach based on network flows modeled by sparse memory networks. In sparse memory networks, we discriminate physical nodes, which represent the systems objects, from state nodes, which describe the dynamics. State nodes are free to represent abstract states and they are not bound to represent, for example, previous steps in memory networks or layers in multilayer networks. We show that various higher-order network flow representations, including memory and multilayer networks, can be represented with sparse memory network.

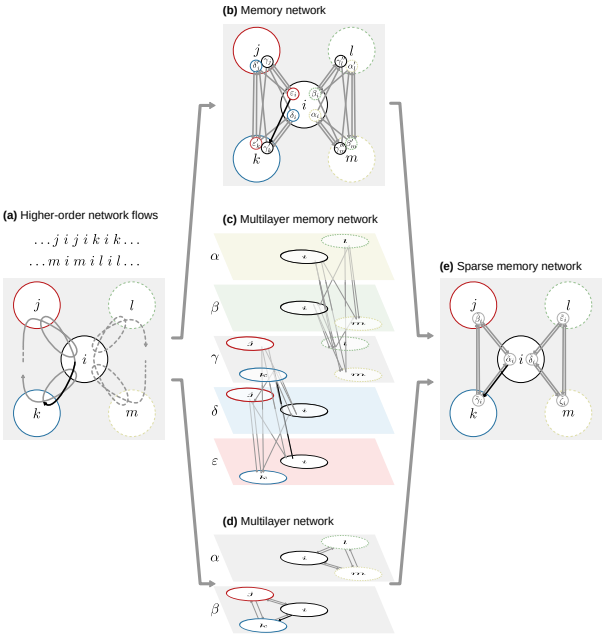


Figure 1: Modeling higher-order network flows with sparse memory networks. (a) Multistep pathways from two sources illustrated on a network with five physical nodes. (b) The pathway data modeled with a second-order Markov model on a memory network. (c) The pathway data modeled on a two-layer network, one layer for each data source. (d) Both memory and multilayer networks mapped on sparse memory network with no redundant nodes. The black link highlights the same step in all representations.

We illustrate our approach with a map of citation flows in science with research fields that overlap in multidisciplinary journals. Compared with currently used categories in science of science studies, the overlapping research fields form better units of analysis because the map more effectively captures how ideas flow through science.

**Corresponding author: ludvig.bohlin@umu.se*

MODELING THE NETWORK DYNAMICS OF PULSE-COUPLED NEURONS

Sarthak Chandra, David Hathcock, Kimberly Crain, Thomas M. Antonsen, Michelle Girvan, Edward Ott

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

A barrier to the computational modeling of neural dynamics in the brain is the practical limit on resources, given the enormous number of neurons in the human brain ($\sim 10^{11}$). Our work addresses this problem by developing a method for obtaining low dimensional macroscopic descriptions for the dynamics of large networks of neurons.

Model

The theta neuron model is a phase oscillator that is used to model the dynamics of Class I excitable neurons. We consider a setup of a large network of such theta neurons, with pulse-like synaptic connections between the neurons as dictated by the network topology. Previous studies modeling the dynamics of such systems have generally been restricted to networks within particular classes of topologies (eg. Ref. [1]). We consider a broad class of network topologies, allowing for arbitrary degree distributions and assortativities (degree correlations).

Methods and Results

We use a mean field approach in conjunction with the analytical techniques of Refs.[2] to study the behavior of pulse coupled theta neurons on networks with arbitrary degree distributions and assortativity, extending results in Ref.[1](fully connected networks), and Ref.[3](networks of Kuramoto oscillators). We analytically obtain a reduced system of equations describing the mean-field dynamics of the system, with a significantly lower dimensionality compared with the complete set of dynamical equations for the system. This dimensional reduction allows for a computationally efficient simulation of the mean-field dynamics of the system, which can be used to calculate an order parameter for the full system, and hence allows for an efficient characterization of phase transitions and attractors. We find that, for sufficiently large well-connected networks, the dynamical behavior of the reduced system agrees well with that of the full network. For networks with tightly peaked degree distributions, the macroscopic behavior closely resembles that of fully connected networks previously studied by others. In contrast, networks with highly skewed degree distributions exhibit different macroscopic dynamics due to the emergence of degree dependent behavior of different oscillators (Fig. 1). In general the long term dynamics of the order parameter can be broadly classified into one of three phases – (1) the partially resting phase; (2) the asynchronously firing phase; and (3) the synchronously firing (SF) phase. We observe that the SF phase of the system can be suppressed by the addition of either assortativity or disassortativity to the network.

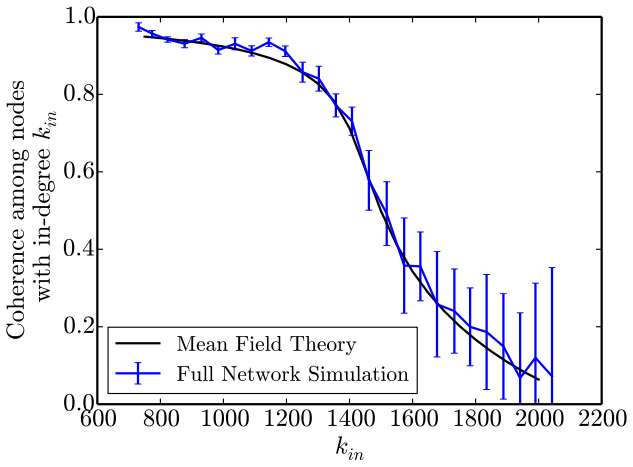


Figure 1: Degree dependent behavior across nodes for a network topology with a highly skewed degree distribution. Note the close agreement between the simulations on the full network and the calculations according to the lower dimensional mean field theory.

References

[1] T. B. Luke, E. Barreto, and P. So. Complete classification of the macroscopic behavior of a heterogeneous network of theta neurons. *Neural Computation*, 25(12):3207–3234, 2013.

[2] E. Ott and T. M. Antonsen. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos*, 18(3):037113, 2008.

[3] J. G. Restrepo and E. Ott. Mean-field theory of assortative networks of phase oscillators. *Europhysics Letters*, 107(6):60006, 2014.

ANALYSING CONSUMER PREFERENCE IN GROCERY STORES USING ANNOTATED NETWORKS

A. Roxana Pamfil, Sam D. Howison, Mason A. Porter

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We quantify customer preference in grocery stores by analysing mesoscale structure in bipartite consumer–product networks. Building on previous work on annotated networks [2], we show that incorporating node metadata into the network model (see Figure 1) reduces the uncertainty in the optimal network partition. Although we focus on consumer behaviour, our methods are applicable to any bipartite network, where analysing the one-mode projections would result in information loss.

Data and motivation

We use anonymised basket-level transaction information to construct a bipartite network of customers and products, where weighted edges correspond to purchases aggregated over a three-month time window. Mesoscale structure in these networks (such as community structure) reveals groups of customers with similar preferences and groups of products that are preferred by different customer types. This information can feed into a system of personalised product recommendations, or it can suggest new segmentations of customers and products, both of which are used by stores for business planning.

For a given network, there may be several different partitions of the nodes into groups that are close to optimal (e.g. that give similarly high likelihood values in a statistical framework). As shown in [2], using additional data in the form of node annotations may improve the quality and interpretability of the optimal partition. For our application, we use product categories as annotations on the product nodes, as illustrated in Figure 1.

Mesoscale structure in annotated networks

We first adapt the algorithm in [2] to bipartite networks; this allows us to find structure in unweighted consumer–product networks. To work with weighted networks instead, we modify the weighted stochastic block model in [1] to incorporate annotations. Although the second approach should be more suitable for our data, we find that the unweighted method consistently gives more meaningful results, likely due to the computational implementation.

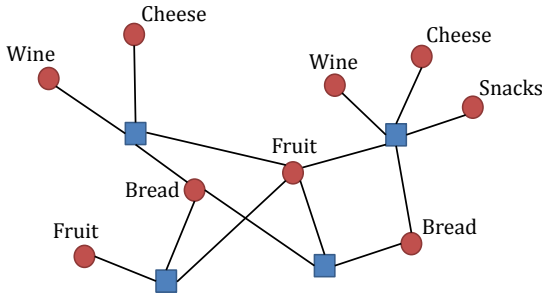


Figure 1: Schematic of a bipartite network of customers (blue squares) and products (red circles). Known product categories serve as annotations on the product nodes.

Results

The algorithm that fits the model to an annotated input network returns a soft partitioning of customer and product nodes into a prespecified number of groups. We find that incorporating annotations tends to increase the probability of a product node belonging to its optimal group, thus reducing the amount of uncertainty in the final partition. This optimal solution reveals sets of customers with distinctive shopping patterns, including a group of price-sensitive customers who prefer snacks and frozen meals and a group of people who are more likely to buy produce, fresh meat, and other ingredients for cooking at home.

References

[1] C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, 3(2):221–248, 2015.

[2] M. E. Newman and A. Clauset. Structure and inference in annotated networks. *Nature Communications*, 7(11863), 2016.

THEORETICAL APPROACH TO POWER GRID ISLANDING

Saleh Soltan[†], *Mihalis Yannakakis*[‡], *Gil Zussman*[†]
[†]*Department Electrical Engineering*, [‡]*Department of Computer Science*
Columbia University, New York, NY
{*saleh,gil*}*@ee.columbia.edu, mihalis@cs.columbia.edu*

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Power Grid Islanding is an effective method to mitigate cascading failures in power grids. The challenge is to partition the power grid network into smaller connected components, called *islands*, so that each island can operate independently for a short period of time. In order for an island to operate, it is necessary that the power supply and demand at that island be almost equal (if the supply and demand are not exactly equal but still relatively close, load shedding/generation curtailing can be used in order for the island to operate). Equality of supply and demand in an island, however, may not be sufficient for its independent operation. It is also important that the infrastructure in that island have the physical capacity to safely carry the power flows. To address this problem, we introduce and study the Doubly Balanced Connected graph Partitioning (DBCP) problem. The DBCP problem is the problem of partitioning a graph into two parts such that both parts are connected and comparable in size, and supply is almost equal to demand in each part. The idea is that when an island is large enough compared to the initial network, it most likely has enough capacity to carry power flows. In this way, the partitions obtained from solutions to the DBCP problem are operational.[1]

Doubly Balanced Connected Graph Partitioning

We introduce and study the Doubly Balanced Connected graph Partitioning (DBCP) problem: Let $G = (V, E)$ be a connected graph with a weight (supply/demand) function $p : V \to \mathbb{Z}$ satisfying $p(V) = \sum_{j \in V} p(j) = 0$. The objective is to partition V into (V_1, V_2) such that $G[V_1]$ and $G[V_2]$ are connected, $|p(V_1)|, |p(V_2)| \leq c_p$, and $\max\{\frac{|V_1|}{|V_2|}, \frac{|V_2|}{|V_1|}\} \leq c_s$, for some constants c_p and c_s. We focus on the case that weights (supply/demand values) are ± 1, but our techniques can be extended, with similar results, to the case in which the weights are arbitrary (not necessarily ± 1), and also to the case that $p(V) \neq 0$ and the excess supply/demand should be split evenly.

The connected partitioning problem with only the size objective has been studied previously. In the most well-known result, Lováz and Gyori [1,3] independently proved that every k-connected graph can be partitioned into k arbitrarily sized connected subgraphs. However, neither of the proofs is constructive, and there are no known polynomial-time algorithms to find such a partition for $k > 3$. The objective of balancing the supply/demand alone, when all $p(i)$ are ± 1, can also be seen as an extension for the objective of balancing the size (which corresponds to $p(i) = 1$).

Since the power grids are designed to withstand a single failure ("$N-1$" standard), and therefore 2-connected, our focus is mainly on the graphs that are at least 2-connected. We use the embedding for k-connected graphs introduced in [2] and show that when G is 2-connected, a solution with $c_p = 1$ and $c_s = 3$ to the DBCP problem always exists and can be found in polynomial time. Moreover, when G is 3-connected, we show that there is always a 'perfect' solution (a partition with $p(V_1) = p(V_2) = 0$ and $|V_1| = |V_2|$, if $|V| \equiv 0 (\text{mod } 4)$), and it can be found in polynomial time.

References

[1] E. Gyori. On division of graphs to connected subgraphs. In *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976)*, volume 1, pages 485–494, 1976.

[2] N. Linial, L. Lovasz, and A. Wigderson. Rubber bands, convex embeddings and graph connectivity. *Combinatorica*, 8(1):91–102, 1988.

[3] L. Lovász. A homology theory for spanning tress of a graph. *Acta Mathematica Hungarica*, 30(3-4):241–251, 1977.

[4] S. Soltan, M. Yannakakis, and G. Zussman. Doubly balanced connected graph partitioning. In *Proc. ACM-SIAM SODA'17*, Jan. 2017.

[1] This abstract summarizes the results that appeared in [4].
This work was supported in part by DTRA grant HDTRA1-13-1-0021, DARPA RADICS under contract #FA-8750-16-C-0054, funding from the U.S. DOE OE as part of the DOE Grid Modernization Initiative, and NSF under grant CCF-1320654 and CCF-1423100.

FEATURE-BASED CLASSIFICATION OF NETWORKS

*Nishant Malik**, *Ian Barnett**, *Marieke L. Kuijjer, Peter Mucha and Jukka-Pekka Onnela*

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Networks are now used to study many systems in scientific and societal domains, where nodes represent system elements and edges represent relationships among these elements. These networks exhibit a broad variety of intricate structural properties. Mathematical measures developed within network science can be used to characterize several structural features of networks. However, this characterization of individual features is unable to provide an automated, statistically principled and computationally efficient method to classify networks in large data sets.

While many network features are common across networks from the same broad class, such as social networks or types of biological networks, we identify finer scales of classifications within such a broad class, leveraging that networks of more similar systems tend to have more similar features. That is, networks representing similar purposes are expected to arise from shared domain specific mechanisms, so it should be possible to classify networks into categories based on features at various structural levels.

We present a novel hybrid approach to network classification, combining manual selection of features of interest with machine learning classifiers. By selecting well-studied features that have been used throughout social network analysis and network science and then classifying with methods such as random forests that are of special utility in the presence of feature collinearity, we find that we achieve higher accuracy, in shorter computation time, with greater interpretability on benchmark problems compared to existing network classification methods. As this hybrid approach relies on a novel combination of existing open-source tools, it can be easily implemented across different application domains to develop classification strategies that are computationally efficient and intuitively understood.

We demonstrate the broad applicability of our approach by classifying days of the week from call detail records, diagnosing types of cancer tumors based on their transcription factor- gene regulatory networks, and testing the method against network classification benchmarks. Furthermore, we extended this approach to classification of recurrence networks used in nonlinear time series analysis and climate networks.

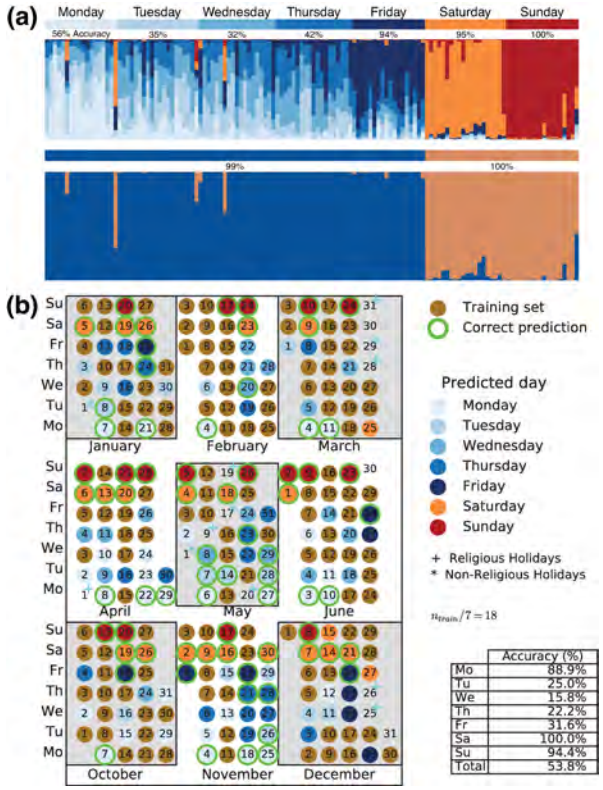


Figure 1: Classification of days of the week from daily communication networks extracted from call record data of a European country. (a) Random forest classification of days of the week: The performance of the 7-day classifier is displayed in the top row with the binary weekend/weekday classifier in the bottom row. (b) KNN classification of days of the week: This visualizes a single realization of classification of days of the week using KNN, where n_{train} is the total number of days used for the training set, which included equal number of days of each day of the week.

*These authors contributed equally to this work.

CONTROLLABILITY IN A NETWORK OF LINEAR DYNAMICAL SYSTEMS

Biswadip Dey, Elizabeth N. Davison, Naomi Ehrich Leonard

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

In studies of the problem of controlling large-scale networks in both network science [1] and control theory [2, 3], the dynamics of individual agents are often assumed to be represented by a scalar integrator. This assumption poses restrictions for many real-world networks where essential behaviors of an agent cannot be captured by modeling it as a single integrator. In this work, we treat the individual agents as multi-input multi-output linear dynamical systems, and then, by letting the interaction between agents be governed by a diffusion process over an undirected graph, we investigate the combined influence of individual agent dynamics and the underlying network topology on the controllability of a networked multi-agent system.

Problem Setup

Here we consider a network of N agents wherein $x_i \in \mathbb{R}^n$, $i = 1, \ldots, N$ represents the state of agent i. By letting $v_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^p$ denote the input and output of agent i, respectively, we define the agent dynamics as

$$\begin{aligned} \dot{x}_i &= Ax_i + Bv_i \\ y_i &= Cx_i, \end{aligned} \tag{1}$$

where $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times p}$ and $C \in \mathbb{R}^{p\times n}$. Moreover, we assume C to be of full-rank. The underlying network topology is encoded by the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \Gamma)$, where each agent corresponds to a node in $\mathcal{V} = \{1, \ldots, N\}$ and $\mathcal{E}$ defines the associated edge set. $\Gamma \in \mathbb{R}_{\geq 0}^{N\times N}$ represents the adjacency matrix associated with $\mathcal{G}$, and γ_{ij}, the elements of Γ, represent the coupling strength along the edge (i, j). The associated graph Laplacian is defined as $L = D - \Gamma$.

In this framework, the input v_i to agent i has two components: (i) a social component governed by diffusive coupling with neighbors, and (ii) an external input (control signal). However, we assume that only m $(1 \leq m \leq N)$ agents have direct access to the external control signal, and these agents constitute the *leader* set $S_m \triangleq \{s_1, \ldots, s_m\}$. Hence, v_i can be expressed as

$$v_i = u_i \mathbf{1}_{S_m}(i) + \sum_{j=1}^{N} \gamma_{ij} C(x_j - x_i), \tag{2}$$

where $\mathbf{1}_{S_m}$ denotes the indicator function of $S_m \subseteq \mathcal{V}$ and $\mathbf{1}_{S_m}(i) = 1$ if $i \in S_m$ and otherwise is 0.

Main Result

By assuming the graph to be undirected and connected, we have $L = L^\intercal \geq 0$. Hence, L can be diagonalized as

$$L = \Psi \Lambda \Psi^\intercal, \tag{3}$$

where $\Psi \in SO(N)$ and $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$, $0 = \lambda_1 < \lambda_2 \leq \cdots \leq \lambda_N$ is the diagonal matrix of eigenvalues of L. This factorization, together with the PBH test [4], leads us to a necessary and sufficient condition for controllability of the overall dynamics (1)-(2).

Theorem 1 *Consider the networked multi-agent system, the dynamics of which are governed by (1) and (2). This system is controllable if and only if the following conditions hold true:*

(I) $[\psi_{s_1,j},\ \psi_{s_2,j},\ \cdots,\psi_{s_m,j}] \neq \mathbf{0}$ *for any* $j = 1, \ldots, N$ *where* ψ_{ij} *are individual elements of the matrix* Ψ*.*

(II) For each $\lambda_i \in spec(L)$*, none of the left eigenvectors of* $(A - \lambda_i BC)$ *are orthogonal to* B*.*

Future Directions

Our result explicitly reveals how the dynamics of individual agents interact with the associated graph Laplacian towards influencing the controllability properties of a given network. In future work, we will characterize sets of leader nodes in such a network of linear dynamical systems that achieve critical control objectives in an optimal way.

References

[1] Y.-Y. Liu and A.-L. Barabási. Control principles of complex systems. *Review of Modern Physics*, 88(3):035006, 2016.

[2] F. Pasqualetti, S. Zampieri, and F. Bullo. Controllability metrics, limitations and algorithms for complex networks. *IEEE Transactions on Control of Network Systems*, 1(1):40–52, 2014.

[3] A. Rahmani, J. Meng, M. Mesbahi, and M. Egerstedt. Controllability of multi-agent systems from a graph-theoretic perspective. *SIAM Journal on Control and Optimization*, 48(1):162–186, 2009.

[4] W. J. Rugh. *Linear System Theory (2nd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1996.

MACROSCOPIC MODELS FOR NETWORKS OF COUPLED BIOLOGICAL OSCILLATORS

Kevin M. Hannay, Daniel B. Forger, Victoria Booth

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We introduce a macroscopic reduction for networks of coupled oscillators motivated by an elegant structure we find in experimental measurements of circadian protein expression and several mathematical models for coupled biological oscillators. The observed structure in the collective amplitude of the oscillator population differs from the well-known Ott-Antonsen ansatz, but its emergence can be characterized through a simple argument depending only on general phase-locking behavior in coupled oscillator systems. We further demonstrate its emergence in networks of noisy heterogeneous oscillators with complex network connectivity. Applying this structure, we derive low-dimensional macroscopic models for oscillator population activity.

Low Dimensional Relations

In large ensembles of coupled phase oscillators we may define the Daido order parameters [3] as,

$$Z_m(t) = R_m(t)e^{i\psi_m(t)} = \frac{1}{N}\sum_{j=1}^{N} e^{im\phi_j(t)}, \qquad (1)$$

where ϕ_j are the phases of the oscillators, R_m are the phase coherences and ψ_m are the mean phases. In 2008, Ott and Antonsen introduced a powerful macroscopic reduction for coupled oscillator networks [1]. In its most powerful form the Ott-Antonsen approach assumes that $R_m = R_1^m, \psi_m = m\psi_1$.

We examined both experimental data [2] and simulations of coupled biological oscillators for a relationship between the Daido order parameters. Surprisingly, we found the Ott-Antonsen relation did not provide a good approximation, however the ansatz,

$$R_m = R_1^{m^2} \qquad \psi_m = m\psi_1 \qquad (2)$$

did describe the systems we tested well.

Emergence

In order to characterize the emergence of our ansatz we consider a model network of N noisy heterogeneous phase oscillators,

$$\dot{\phi}_i = \omega_i + \frac{K}{d_i}\sum_{j=1}^{N} A_{ij}H(\phi_j - \phi_i) + \sqrt{D}\eta_i(t), \qquad (3)$$

where η_i is a white noise process. Numerical simulations of Equation. 3 with a Gaussian distribution of frequencies (ω_i) shows that our ansatz robustly emerges in the network as the coupling strength increases (Fig. 1). We give a simple analytical argument which explains the prevalence of our ansatz in biological networks.

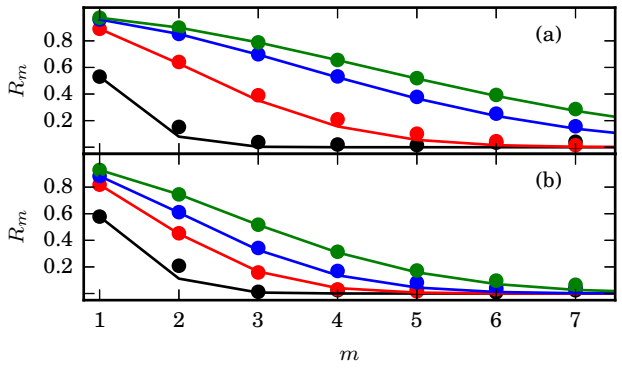


Figure 1: (a) Barabasi-Albert Scale-Free network (b) Watts-Strogatz Small World network. Circles show the results from simulations of networks (Eq. 3). Solid lines show $R_m = R_1^{m^2}$. Colors differetiate coupling strengths.

Macroscopic Model

Finally, we demonstrate how our ansatz may be used to extract low-dimensional macroscopic models for networks of coupled biological oscillators.

References

[1] Ott, E. & Antonsen, T. M. Low dimensional behavior of large systems of globally coupled oscillators. *Chaos* **18**, 037113 (2008).

[2] Abel, J. H. *et al.* Functional network inference of the suprachiasmatic nucleus. *Proc. Natl. Acad. Sci.* **113**, 4512–4517 (2016).

[3] Daido, H. Critical conditions of macroscopic mutual entrainment in uniformly coupled limit-cycle oscillators. *Prog. Theor. Phys.* **89**, 929–934 (1993).

GENERALIZED HYPERGEOMETRIC ENSEMBLES: STATISTICAL HYPOTHESIS TESTING IN COMPLEX NETWORKS

Giona Casiraghi, Vahan Nanumyan, Ingo Scholtes and Frank Schweitzer

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

The statistical analysis of networks requires analytically tractable probability distributions that quantify the probability to observe a network under a null hypothesis. In this talk we will present the class of *generalized hypergeometric ensembles*, which provides a powerful framework to perform statistical hypothesis testing and model selection in complex networks.

Abstract

Statistical ensembles of networks, i.e., probability spaces of all networks that are consistent with given aggregate statistics, have become instrumental in the analysis of complex networks [1]. Their numerical and analytical study provides the foundation for the inference of topological patterns [5, 3], the definition of network-analytic measures [5], as well as for model selection and statistical hypothesis testing [2]. Contributing to the foundation of these data analysis techniques, we introduce *generalized hypergeometric ensembles* (gHypEs), a broad class of analytically tractable statistical ensembles of finite, directed and weighted networks.

This framework is a generalization of the classical configuration model [4], commonly used to randomly generate networks with given degree sequence or distribution. Different from this, we utilize an *edge-centric sampling* of m edges from the set of all possible edges, such that the sequence of *expected* degrees of nodes is preserved. For each pair i, j of the n nodes, we sample edges from a set of Ξ_{ij} possible multi-edges uniformly at random. This can be viewed as an *urn problem* where edges to be sampled are represented by balls in an urn. We specifically obtain an urn with $M = \sum_{i,j} \Xi_{ij}$ balls having $n^2 = |V \times V|$ different colours, representing all possible edges between a given pair of nodes. Each adjacency matrix $\mathbf{A}$, with entries A_{ij} such that $\sum_{i,j} A_{ij} = m$, corresponds to one particular realization drawn from this ensemble. The probability to draw exactly $\mathbf{A} = \{A_{ij}\}_{i,j \in V}$ edges between each pair of nodes is given by the multivariate hypergeometric distribution. Moreover, by biasing the above mentioned sampling, we can further generalise the ensemble such that each pair of nodes has a given *propensity* to form an edge, i.e. arbitrary *degree-corrected* tendencies of pairs of nodes to form edges between each other. This new sampling process is described by a *biased urn*, whose sampling probability is the multivariate non-central Wallenius hypergeometric distribution [6].

Studying empirical and synthetic data, we show that this class of ensembles provides a powerful framework for model selection and hypothesis testing in complex networks. We demonstrate how gHypEs can be used to develop statistical regression models to analyze data in the form of complex networks. The resulting non-linear parametric models take as independent variables diverse hypotheses about the network structure, e.g. community membership or more complicated node-node relations. They allow then to regress the influence of such hypotheses on the network topology, estimating the intensity and the significance of their effects. The goodness of fit of a regression model can be assessed by means of likelihood-ratio tests, or by computing the mahalanobis distance of the network according to the chosen model.

References

[1] Ginestra Bianconi. Entropy of network ensembles. *Phys. Rev. E*, 79:036114, Mar 2009.

[2] Paul W Holland and Samuel Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50, 1981.

[3] Brian Karrer and M. E. J. Newman. Stochastic blockmodels and community structure in networks. *Phys. Rev. E*, 83:016107, Jan 2011.

[4] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random Structures & Algorithms*, 6(2-3):161–180, 1995.

[5] Mark E. J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, jun 2006.

[6] Kenneth T Wallenius. *Biased Sampling: the Noncentral Hypergeometric Probability Distribution*. Ph.d. thesis, Stanford University, 1963.

COMPRESSING OVER-THE-COUNTER MARKETS

Tarik Roukny (Massachusetts Institute of Technology), Marco D'Errico (University of Zurich)

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

Over-the-counter (OTC) derivative markets were regarded as one of the key factors contributing to the Global Financial Crisis of 2007-2008, which has had long-standing societal implications. Unlike centrally organized markets, participants in OTC markets trade on a bilateral basis, engendering large networks of contractual obligations and risk transfers. These markets are also known to be opaque (i.e., market information is often very limited to most agents) and large in size (the aggregate volumes of total bilateral obligations can amount to several trillion dollars [2, 1]). The size, coupled with the lack of transparency of these markets has become an important concern for policy makers [3].

In this paper, we show both theoretically and empirically that the size and complexity of OTC markets can be reduced without affecting individual trade balances. First, we find that the networked nature of these markets generates an excess of obligations: a significant share of the total market volume can be deemed redundant. Second, we show conditions under which such excess can be removed while preserving individual net positions. We refer to this netting operation as compression and identify feasibility and efficiency criteria, highlighting intermediation as the key element for excess levels. We show that a trade-off exists between the amount of excess that can be eliminated from markets and the conservation of trading relationships. We then design several compression benchmark solutions and test their efficiency using a unique and granular dataset on credit-default swap transactions involving all EU firms and their global counterparties. The compression benchmarks gradually differ in their capacity to modify the markets initial web of outstanding trades. We find that, between 2014 and 2016, on average more than 75% of the total notional in markets eligible for compression. While bilateral compression is shown to be limited, more sophisticated compression techniques, which can identify longer chains of compressible contracts, generally remove most of the eligible outstanding notional. In particular, even the most conservative multilateral compression approach, which does not alter trading relationships, reaches up to 98% of notional elimination.

While some markets have already adopted compression in order to reduce their risk and size, these results show, for the first time, the efficiency and trade-offs of compression when systematically applied at a larger scale. Finally, our framework provides ways for regulators and policymakers to curb the impact of financial crises and improve the efficiency of markets by reducing the total aggregate size of markets and reconfiguring the web of obligations.

Keywords

OTC markets, compression, intermediation, financial network

References

[1] J. Abad, I. Aldasoro, C. Aymanns, M. D'Errico, L. Fache Rousova, P. Hoffmann, S. Langfield, M. Neychev, and T. Roukny. Shedding light on dark markets: first insights from the new EU-wide OTC derivative dataset. *ESRB Occasional Paper Series No 11-16*, 2016.

[2] BIS. OTC derivatives statistics at end-June 2015. Technical report, Basel Committee on Banking Supervision - Bank of International Settlements, 2015.

[3] D. Duffie. *Dark markets: Asset pricing and information transmission in over-the-counter markets*. Princeton University Press, 2012.

EIGENVECTOR-BASED CENTRALITY MEASURES FOR TEMPORAL NETWORKS

Dane Taylor, Sean A. Myers, Aaron Clauset, Mason A. Porter and Peter J. Mucha

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Numerous centrality measures have been developed to quantify the importances of nodes in time-independent networks, and many of them can be expressed as the leading eigenvector of some *centrality matrix* $\mathbf{C}$. Because many networks depend on time, we introduce (see [1]) a principled generalization of centrality measures that is valid for any eigenvector-based centrality. We consider a temporal network with N nodes as a sequence of T layers that describe the network during different time windows, and we couple centrality matrices $\{\mathbf{C}^{(t)}\}$ for the layers $t \in \{1, \ldots, T\}$ into a *supra-centrality matrix*,

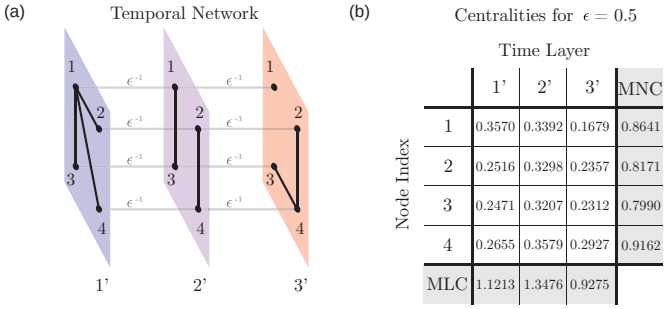
$$\mathbb{C}(\epsilon) = \begin{bmatrix} \mathbf{C}^{(1)} & \epsilon^{-1}\mathbf{I} & 0 & \cdots \\ \epsilon^{-1}\mathbf{I} & \mathbf{C}^{(2)} & \epsilon^{-1}\mathbf{I} & \ddots \\ 0 & \epsilon^{-1}\mathbf{I} & \mathbf{C}^{(3)} & \ddots \\ \vdots & \ddots & \ddots & \ddots \end{bmatrix}. \qquad (1)$$

Coupling strength $\epsilon > 0$ is a tuning parameter that controls the rate at which centralities change over time.

We study $\mathbb{C}(\epsilon)$ for synthetic and empirical network datasets, and our work consists of two main parts: (i) interpreting the length-NT dominant eigenvector of $\mathbb{C}(\epsilon)$ for centrality analysis, and (ii) conducting a singular perturbation analysis for the *time averaging* limit $\epsilon \to 0^+$.

Joint, Marginal and Conditional Centralities

Each entry $\mathbb{v}_j(\epsilon)$ in the dominant eigenvector of $\mathbb{C}(\epsilon)$ encodes a "joint centrality" that reflects both the importance of physical node $i = \text{mod}(j, N)$ and time layer $t = \lceil j/N \rceil$. It is convenient to represent the length-NT eigenvector $\mathbb{v}(\epsilon)$ by an $N \times T$ matrix so that entry $W_{it} = \mathbb{v}_{i+N(t-1)}(\epsilon)$ encodes the joint centrality of physical node i at time t. In Fig. 1(a), we illustrate an example synthetic network with $T = 3$ and $N = 4$. We indicate the associated joint centralities for $\epsilon = 0.5$ with a table in Fig. 1(b). The shaded regions in Fig. 1(b) describe two new concepts that we call "marginal node centralities" (MNC) $\{x_i = \sum_t W_{it}\}$ and "marginal layer centralities" (MLC) $\{y_t = \sum_i W_{it}\}$, which provide *uncoupled* centralities. We study how the centrality of each node evolves over time by also defining "conditional node centralities" $\{Z_{it} = W_{ij}/y_t\}$, which quantify the importances of physical nodes at time t relative to other physical nodes at that particular time.



	1'	2'	3'	MNC
1	0.3570	0.3392	0.1679	0.8641
2	0.2516	0.3298	0.2357	0.8171
3	0.2471	0.3207	0.2312	0.7990
4	0.2655	0.3579	0.2927	0.9162
MLC	1.1213	1.3476	0.9275	

Figure 1: *Example temporal network.* (a) Intra-layer edges (black lines) encode the network at different instances and inter-layer identity edges (gray lines) couple these layers together with strength ϵ^{-1}. (b) For coupling strength $\epsilon = 0.5$, we indicate joint node-layer centralities $\{W_{it}\}$ (white boxes), MNC $\{x_i\}$ (right-most row) and MLC $\{y_t\}$ (bottom row).

Time-Averaged Centrality and First-Order-Mover Scores

Joint, marginal and conditional centralities depend on coupling strength ϵ, which tunes the rate at which the nodes' centralities change over time. We conduct a singular perturbation analysis for the limit $\epsilon \to 0^+$, which implements a time averaging in that the conditional node centralities become constant in time, $Z_{it} \to \alpha_i$ for every t. We refer to the values $\{\alpha_i\}$ as the nodes' "time-averaged centralities," and we find that they correspond to the zeroth-order terms of a singular perturbation expansion. Interestingly, the values $\{\alpha_i\}$ correspond to the dominant eigenvector for a weighted average of centrality matrices, $\sum_t w_t \mathbf{C}^{(t)}$. We also study first-order terms to define "first-order-mover scores" that concisely describe the magnitude to which nodes' centralities change over time.

References

[1] D. Taylor, S.A. Meyers, A. Clauset, M.A. Porter and P.J. Mucha, *Multiscale Modeling and Simulation* 15(1), 537–574 (2017).

ASYMMETRY-INDUCED SYNCHRONIZATION IN MULTILAYER NETWORKS

Yuanzhao Zhang, Takashi Nishikawa and Adilson E. Motter

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

A scenario has recently been reported in which in order to stabilize complete synchronization of an oscillator network—a symmetric state—the symmetry of the system itself has to be broken by making the oscillators nonidentical. But how often does such behavior—which we term asymmetry-induced synchronization (*AISync*)—occur? Here we present a general scheme for constructing *AISync* systems and demonstrate that this behavior is the norm rather than the exception in a wide class of physical systems that can be seen as multilayer networks. This framework doesn't depend on specific nodal dynamics and provides novel insights into the phenomenon of *AISync*.

Introduction

A general belief in the field of network dynamics is that homogeneity in the local dynamics and interaction network can facilitate complete synchronization. It has been shown, however, that structural heterogeneity in networks of identical oscillators or oscillator heterogeneity in structurally symmetric networks can stabilize otherwise unstable synchronous states, thus effectively breaking the symmetry of a system to stabilize a symmetric state. These scenarios can be interpreted as the converse of symmetry breaking, and hence also a converse of chimera states.

Results

We first identify the class of all structurally symmetric networks by generalizing the vertex-transitive graphs (in which each node can be mapped to any other node through node permutations that leave the network invariant) from algebraic graph theory to directed edges and multiple edge types. These are the ideal networks for the study of *AISync* due to their structural homogeneity and fundamental role in cluster synchronization (as symmetry clusters).

Consider a symmetric network of N (not necessarily identical) oscillators coupled through K different types of interactions [e.g., Fig. 1(a)]. Determining the stability of such systems is extremely challenging, and even the existence of a synchronous state is not guaranteed. In this work we introduce and characterize a wide class of *multilayer systems* of nonidentical oscillators, whose stability can nonetheless be analyzed by the master stability formalism (MSF) and exhibit *AISync* generically.

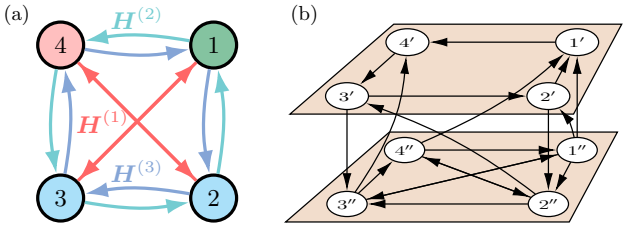


Figure 1: Multilayer construction of *AISync* networks. (a) Example of a symmetric network of $N = 4$ heterogeneous oscillators and $K = 3$ types of (directed) links with different interaction functions. (b) One of many possible multilayer networks corresponding to the network in (a), with $L = 2$ layers and $n = LN = 8$ identical subnodes.

In our multilayer system, each node can be further decomposed into L identical *subnodes*, belonging to L different layers and interacting through a set of *internal sublinks*. The pattern of these internal sublinks is thus part of the node's properties and can be used to represent node heterogeneity. For a pair of connected nodes, the type of the connecting link is determined by the pattern of *external sublinks* between the subnodes of these two nodes (see Fig. 1(b) for an $L = 2$ example). Since subnodes and sublinks are identical, we can directly apply the MSF analysis by flattening the multilayer network into a monolayer network of diffusively coupled subnodes.

We have discovered an abundance of concrete examples of *AISync* systems under this multilayer framework, including those with periodic and chaotic dynamics, directed and undirected coupling schemes, and continuous- and discrete-time dynamics. In particular, experimentally testable *AISync* systems are constructed using optoelectronic networks. Finally, since a symmetric network in complete synchrony is the basic building block of cluster synchronization in more general networks, *AISync* should be common also in facilitating cluster synchronization by breaking the symmetry of the cluster subnetworks.

RANKING IN GRAPHS BY NUMERICALLY APPROXIMATING KATZ CENTRALITY

Eisha Nathan, Geoff Sanders, David A. Bader

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

Identifying highly ranked vertices in graphs is a common query using centrality measures (specifically Katz Centrality). We use iterative solvers to obtain an approximate solution to a ranking vector on the vertices of the graph and use the residual to accurately certify how much of the approximation matches the unknown exact solution.

Introduction

Katz centrality rankings quantify the ability of a vertex to initiate walks in the graph, while penalizing long walks by a fixed factor α. Calculating Katz scores exactly is prohibitively computationally expensive ($\mathcal{O}(n^3)$) so in practice iterative methods are often used to obtain an approximation. We prove that the differences between the approximation and unknown exact solution guarantee how far down the ranking we can go before the approximation error makes it unreliable.

Definitions and Theory

For a graph $G = (V, E)$ with V the set of n vertices and E the set of m edges with adjacency matrix A, we define the Katz centrality of vertices in the graph as the $n \times 1$ vector $\mathbf{c}^* = A(I - \alpha A)^{-1}\mathbf{1}$ [1].

Table 1: Notation used in this paper.

Name	Definition
A	Adjacency matrix, $a_{ij} = 1$ if $(i,j) \in E, 0$ else
M	$I - \alpha A$
$\mathbf{x}^*, \mathbf{c}^*$	Exact solutions, $\mathbf{x}^* = M^{-1}\mathbf{1}, \mathbf{c}^* = A\mathbf{x}^*$
$\mathbf{x}^{(k)}, \mathbf{c}^{(k)}$	kth approximations (from iterative solver) to $\mathbf{x}^*, \mathbf{c}^*$
r_k	Residual, $\|\mathbf{1} - M\mathbf{x}^{(k)}\|_2$
$\lambda_{min}(M)$	Smallest eigenvalue of M

Theorem 1. *If* $|c_i^{(k)} - c_j^{(k)}| > 2\epsilon_k$ *for* $\epsilon_k = \frac{\|A\|_2}{\lambda_{min}(M)} r_k$, *then the ranking of vertex* i *above* j *is correct.*

Proof. We can bound the point-wise error in the ranking to provide a necessary gap to certify correctness of elements in the approximation.

$$\begin{aligned}\|\mathbf{c}^* - \mathbf{c}^{(k)}\|_\infty &\leq \|\mathbf{c}^* - \mathbf{c}^{(k)}\|_2 = \|A\mathbf{x}^* - A\mathbf{x}^{(k)}\|_2 \\ &\leq \|A\|_2\|\mathbf{x}^* - \mathbf{x}^{(k)}\|_2 = \|A\|_2\|M^{-1}\mathbf{1} - \mathbf{x}^{(k)}\|_2 \\ &\leq \|A\|_2\|M^{-1}\|_2 r_k \leq \frac{\|A\|_2}{\lambda_{min}(M)} r_k =: \epsilon_k\end{aligned}$$

Since $c_i^{(k)} - c_i^* < \epsilon_k$ and $c_j^* - c_j^{(k)} < \epsilon_k$, if $c_i^{(k)} - c_j^{(k)} > 2\epsilon_k$, then $c_i^* - c_j^* > 0$ and rank$(i) >$ rank(j) is correct. □

Experiments

Currently to identify top vertices, we run an iterative solver to machine precision ($\approx 10^{-15}$). We develop a new stopping criterion to find top vertices with previously missing theoretical guarantee of correctness. For top R vertices with desired precision $\phi_0 \in (0, 1]$: terminate solver when $|c_R^{(k)} - c_j^{(k)}| > 2\epsilon_k$, if current precision $\frac{R}{j-1} > \phi_0$. Figure 1 shows the reduction in # iterations running to machine precision (I_E) vs. # iterations with the new stopping criterion (I_A). Using 38 real-world graphs and conjugate gradient as the iterative solver, we obtain an average of 4.54× reduction with a maximum of 19.6×, which is significant because running to machine precision can take up to 1000s of iterations.

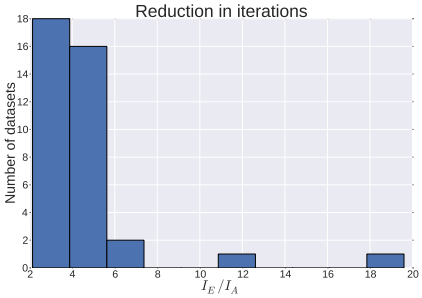


Figure 1: R=100, ϕ_0=0.95.

Conclusions and Future Work

We bridge the two fields of numerical analysis and network analysis by understanding how the error in a linear solver affects the data analysis problem of ranking. By bounding the error in an approximate solution from an iterative method, we can identify the most central vertices with high confidence. Future work will study the theoretical guarantees in a personalized setting, if we only desire the Katz scores averaged from a user-desired set of seed vertices, and extend our theory to directed graphs.

Acknowledgments

This work was partially performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344 with release number LLNL-ABS-732732.

References

[1] L. Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

PROPAGATION OF CASCADING OVERLOAD FAILURES IN INTERCONNECTED NETWORKS

Malgorzata Turalska, Ananthram Swami

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Cascading failures are frequently observed in networked systems and remain a major threat to the reliability of network-like infrastructure. To assess system resilience, we analyze the effect of link failure on the process of the sandpile avalanche propagation through interconnected networks. We observe a positive feedback between link failure due to overuse and sandpile dynamics, where damage spread is controlled by the link strength and density of interlayer connections. Our work provides insight into the problem of optimal robustness of systems of interconnected networks.

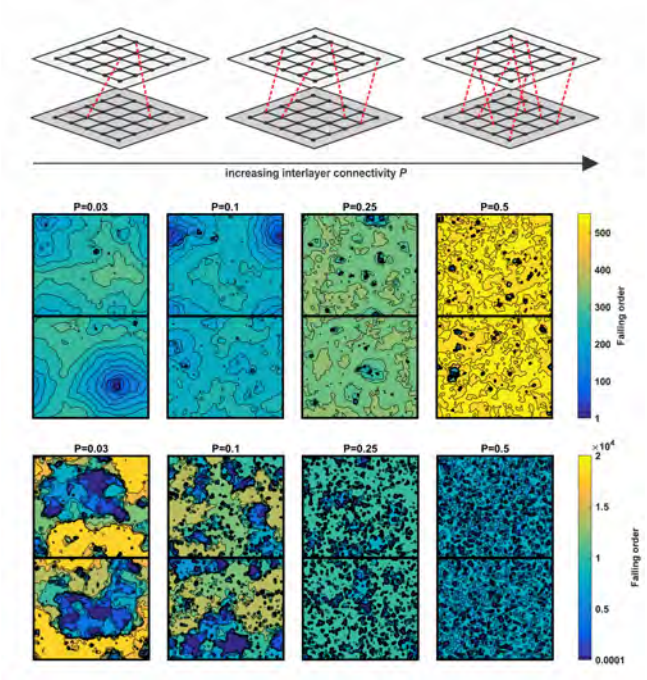


Figure 1: Link failing order for low *(top)* and high *(bottom)* usage threshold θ as a function of increasing interlayer connectivity P. For low θ and P damage spreads in a wave-like manner, while high θ leads to gradual fragmentation of the system. Increased coupling between layers results in further fragmentation as damage spread originates from more sites simultaneously. Adjacent color maps correspond to the behavior of two layers of the system.

Model of overload failures

Here we consider a classic model of cascading failure, the BTW sandpile model, on a system of interdependent networks. Additionally we assume that links in the system fail after they have transported more than θ grains of sand. For simplicity and ease of visualization, we consider a system of two square lattices, with periodic boundary conditions in each layer, where both inner and interlayer links are characterized by the same strength θ.

Propagation of overload failures

In a weakly connected system where θ is low structural damage to the network propagates radially from a site of initial failure causing an abrupt collapse of the entire system (Fig.1; *top, left*). An increase of link strength θ causes more gradual and uncorrelated damage spread, with different parts of the system failing at different times (Fig.1; *bottom, left*). In both cases an increase in coupling P between layers leads to increase in the number of sites at which failures originate followed by simultaneous destruction of remaining links (Fig.1; *right*).

Strong and weak links, however, affect system resilience in diametrically different manner. Increase of coupling P between layers in a system with weak links leads to greater diversity of times at which links fail, with an abrupt collapse of the network occurring at later times (Fig.1; *top, right*). Thus when operating a system built on weak components the increase of coupling between layers comes as a strategy improving resilience to failures. On the other hand, an optimal resilience for a system of strong components is reached at low connectivity, where greater variability of failing times is observed.

These results come in line with observations of numerous nature and man-made networks characterized by a modular structure, where clusters of strongly connected nodes are weakly coupled with each other. Our work suggests that such mixed topology might be most robust one with respect to failures propagating through the system.

HIGHER-ORDER CLUSTERING COEFFICIENTS

Austin R. Benson, Hao Yin, Jure Leskovec, David F. Gleich

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

The tendency for real-world networks to cluster is the basis for many models and algorithms for complex networks. The standard measurement of this tendency is the clustering coefficient, which is the probability that a length-2 path is "closed", i.e., induces a triangle. However, higher-order structures beyond triangles are crucial to understanding complex networks, and the clustering behavior with respect to such structures is not well understood. Here we introduce higher-order clustering coefficients, which measure the closure probability of higher-order cliques and reveal new insights into how networks cluster.

Generalizing the clustering coefficient and data insights

In many domains, clustering comes from temporal closure patterns, particularly through the closure of a length-2 path into a triangle. This pattern, commonly referred to as *triadic closure*, occurs throughout social and information networks [1, 4]. In other cases such as metabolic networks, clustering arises from dense modules operating within a larger system [2]. In general, clustering is a tendency for lower-order structures (e.g., edges) to form higher-order structures (e.g., triangles or dense modules).

The prototypical measurement for the extent to which the nodes of a network form clusters is the *clustering coefficient*, which is the fraction of length-2 paths that induce a triangle [3]. However, the clustering coefficient is inherently restrictive as it measures the closure pattern of just one simple structure—the triangle, or 3-clique. In this work, we generalize the clustering coefficient to account for higher-order closure patterns.

Our generalization is based on an alternative interpretation of the clustering coefficient as a form of clique expansion. Specifically, consider any 2-clique K in a network (that is, a single edge). Now "expand" K by attaching an edge e adjacent to K (i.e., e and K share exactly one node). The clustering coefficient C is then the fraction of (2-clique, adjacent edge) pairs that are *closed*, meaning that the pair induces a $(2+1)$-clique, or a triangle. For higher-order clustering coefficients, instead of expanding 2-cliques to 3-cliques, we simply expand ℓ-cliques to $(\ell+1)$-cliques (Fig. 1). This definition maintains many nice properties of the classical clustering coefficient such as global and local network measurements, probabilistic interpretations, and computational feasibility.

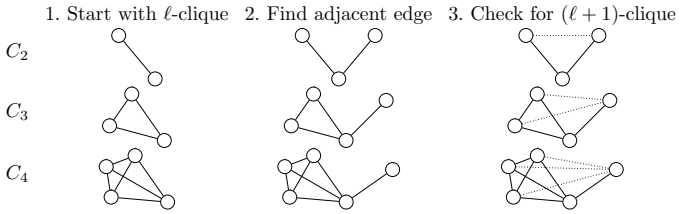


Figure 1: Our ℓth-order clustering coefficient C_ℓ is the probability that an (ℓ-clique, adjacent edge) pair is closed, i.e., induces an $(\ell+1)$-clique.

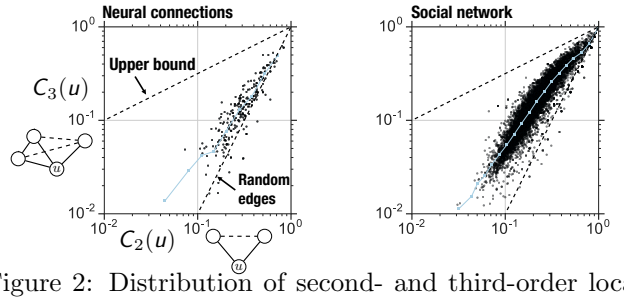


Figure 2: Distribution of second- and third-order local clustering coefficients over all nodes in two networks.

In addition to developing mathematical properties of higher-order clustering coefficients, we use them to gain insights into real-world networks. For example, while nodes in both a neural and a social network have high local clustering in the traditional sense, only nodes in the social network network exhibit higher-order clustering (Fig. 2). We also show that a network with large higher-order clustering must have a 1-hop neighborhood with small clique-based conductance. We use this to find good seeds for local clustering with personalized PageRank.

References

[1] E. M. Jin, M. Girvan, and M. E. Newman. Structure of growing social networks. *Physical Rev. E*, 2001.
[2] E. Ravasz and A.-L. Barabási. Hierarchical organization in complex networks. *Physical Rev. E*, 2003.
[3] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 1998.
[4] Z.-X. Wu and P. Holme. Modeling scientific-citation patterns and other triangle-rich acyclic networks. *Physical Rev. E*, 2009.

TEMPORAL-STRUCTURE-PRESERVING NETWORK TRANSFORMATIONS FOR CHARACTERIZING INFORMATION SPREADING CAPACITY

Mingwu Li, Vikyath D. Rao, Tim Gernat, Harry Dankowicz

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We propose two novel structure-preserving network transformations and associated null models for the study of information spreading on temporal networks. Both transformations build on the commonly used randomly-permuted-times null model (PT) which randomizes contact timestamps on edges. By preserving the lifespan of edges (PTE) or nodes (PTN), our null models eliminate artifacts introduced by the PT model when the ongoing link picture does not hold. We use the proposed transformations to estimate the information spreading capacity i) of synthetic networks with heterogeneous edge lifespans, and ii) empirical networks with nodes entering and leaving dynamically. Our analysis shows that predictions of spreading capacity change significantly with the choice of null model, putting in question earlier results based solely on PT null models.

Temporal networks and randomized null models

Time-dependent networks exhibit different temporal structures, e.g., bursty interevent times (IETs), bursty edge activation dynamics, and heterogeneous edge and node lifespans. It is natural to study the influence of an observed structure on the information spreading capacity of a network by comparing measurements of simulated spreading on the network against those in an appropriately constructed null model that destroys this structure.

The randomly-permuted-times null model (PT), which randomizes contact timestamps on edges, is commonly used in the literature to study the effects of a bursty IET distribution [2]. In addition to destroying burstiness, however, PT also changes several other temporal structures, e.g., mean IET as well as edge and node lifespans. This can introduce artifacts in the randomized reference network that unintentionally bias its spreading properties. For instance, an increase in the active lifespan of nodes may inaccurately reflect the network behavior in cases where nodes are dynamically entering and leaving the network. To address these limitations, we introduce two new null models that preserve the first and last timestamps of edges (PTE) or nodes (PTN), while permuting timestamps of contacts that occurred within the lifespan of edges/nodes.

Simulated spreading on temporal networks

To study the effects of temporal structure on the spreading capacity of a network, we perform simulations of deterministic susceptible-infected (SI) dynamics. For each network, we run 500 simulations with random initial infections and extract mean prevalence curves from the data. We use the time to reach 20% prevalence to characterize the speed-up and slow-down of spreading in each original network relative to the corresponding null models.

We study synthetic networks (using the method in [1] albeit without rescaling in order to control IET statistics) with varying heterogeneity in edge lifespans to highlight the difference between the PT and PTE null models. We use empirical networks, e.g., the sexual contact network in [3], to investigate the difference between PT and PTN null models when nodes enter and leave dynamically.

Results and discussion

Our analysis shows that application of the PT network transformation may lead to the incorrect conclusion that networks such as the sexual contact network show a large speed-up in spreading, whereas this does not follow when preserving node lifespans as ensured by the PTN transformation. Similarly, we find that differences in predicted spreading on synthetic networks based on PTE and PT null models, respectively, become significant with increased heterogeneity in edge lifespans. Finally, comparison of spreading dynamics on empirical and synthetic networks and the corresponding PTE null models suggest that bursty edge IETs, in fact, slow down spreading.

References

[1] P. Holme. Epidemiologically optimal static networks from temporal network data. *PLoS Comput Biol*, 9(7):e1003142, 2013.

[2] P. Holme and J. Saramäki. Temporal networks. *Physics reports*, 519(3):97–125, 2012.

[3] L. E. Rocha, F. Liljeros, and P. Holme. Simulated epidemics in an empirical spatiotemporal network of 50,185 sexual contacts. *PLoS Comput Biol*, 7(3):e1001109, 2011.

A FRAMEWORK FOR CASCADE SIZE CALCULATIONS ON RANDOM NETWORKS

Rebekka Burkholz, Frank Schweitzer

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We develop a branching process approximation of the cascade size evolution for a large class of models on infinitely large (configuration model type) random networks [1]. Our approach allows us to identify a basic trade-off between cascade exposure and damage diversification that determines whether cascades kick-off.

Branching process approximations of the cascade size

Cascade processes are a widely observed phenomenon in an interdependent world. Already the failure of a few components of a system can trigger the failure of dependent components and set off a cascade of successive failures that endanger the functioning of the system as a whole. Examples include financial institutions that face insolvency, fiber bundles that break under stress, or traffic nodes that distribute overload in case of congestion.

Many theoretical investigations of such phenomena are concerned with the question how the network topology and robustness of nodes contribute to the risk of large cascades. As proxy, usually the average final fraction of failed nodes ρ in random graph ensembles with given degree distribution is analyzed. For some models, ρ can be iteratively calculated in the (thermodynamic) limit of infinitely large network size with the help of a branching process approximation, also known as local tree or heterogeneous mean field approximation. In comparison with Monte Carlo simulations, these calculations usually save considerable computational time and efforts, while they further deepen the theoretical understanding of the key factors driving a cascade.

Commonly, they involve compositions of generating functions corresponding to discrete probability distributions, for instance, the degree distribution of a network. However, this approach breaks down when potentially continuous and heterogeneous distributions determine the dynamics of a process. To overcome this obstacle, we present an alternative view on branching process approximations and shift the perspective towards the iterative update of suitable probability distributions. Within this framework, we are able to correctly compute the whole time evolution of the average cascade size for a large class of cascade processes. This allows to consider the recovery of nodes and interventions at specific times also analytically. Further, our approach captures certain fiber bundle and overload redistribution models that could not be tackled before. The key novelty is the introduction of a random variable L that describes the impact that a node can possibly have on its network neighbors, when it has failed at some point before a considered point in time. This variable carries all necessary information about the former time steps and respects the Markovian nature of the studied cascade processes.

This way, we can compare several cascade processes involving a form of load distribution mechanism in case of a node's failure. We encounter a basic trade-off between two effects that are fostered by the presence of hubs and and overall high system connectivity: damage diversification and cascade exposure. The failure time of hubs is essential in deciding which effect outweighs the other. To see this, let's assume that a failing node splits a certain amount of load between its (functional) network neighbors and this load causes a damage that increases with the amount of distributed load. On the one side, a failing node inflicts a lower damage to each of its neighbors if it has a higher degree, i.e. the load is shared between a higher number of neighbors. On the other side, it inflicts damage also to a higher number of nodes in the network. Thus, its failure has also the potential to trigger considerable further failures. Furthermore, a high degree node itself is exposed to a high number of potential load distributing neighbors. So, it might face an increased failure risk. However, the early failure of hubs can also prevent the cascade amplification by disconnecting parts of the system and thus blocking possible cascade paths. Additionally, it can hinder the accumulation of high loads that would eventually be distributed at a later point in time.

References

[1] R. Burkholz and F. Schweitzer. A framework for cascade size calculations on random networks. *arXiv preprint*, 2017.

DATA-DRIVEN MODELS OF BRAIN NETWORK DYNAMICS PREDICT INDIVIDUAL DIFFERENCES IN PERFORMANCE ON COGNITIVELY DEMANDING TASKS

Kanika Bansal, John D. Medaglia, Danielle S. Bassett, Jean M. Vettel and Sarah F. Muldoon

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Humans show significant variability in cognitive task performance, and the origin of this variability is not well understood. Using a data-driven computational model of human brain dynamics, we demonstrate that the underlying structural organization of individual brain networks accounts for a significant portion of variability in three language tasks.

Description

Diffusion spectrum imaging (DSI) provides information about the structural (anatomical) connectivity between brain regions. This data serves as a fundamental basis for efforts aimed at enhancing our understanding of the organization of the human brain as a complex and efficient network. How important is this basic anatomical skeleton of the brain in explaining and predicting individual differences in cognition? To address this question, we use a computational model that combines data-driven structural connectivity with nonlinear Wilson-Cowan oscillators [1, 2] to study the spatiotemporal dynamics of a human brain. We study simulated functional activity both within the global brain network and throughout task specific sub-networks across a cohort of individuals. We then construct functional measures to explain individual performance across three different language tasks. We find that task performance correlates with the activation of either local or global circuitry depending on the complexity of the task. Motivated by experimental data, we also stimulate the left inferior frontal gyrus of the model. We quantify the spread of the stimulation and then use the patterns of activation to explain the effect of stimulation on task performance. By emphasizing differences in underlying structural connectivity, our model is a powerful tool to differentiate and predict individual performance on tasks that vary in complexity.

Computational model

The anatomical structure of an individual's brain is represented as a network whose connectivity is obtained from the density of streamlines connecting different brain regions (network nodes) as determined from DSI data. The dynamics of each brain region is modeled by a nonlinear Wilson-Cowan oscillator [2] where the average firing rate of excitatory (E) and inhibitory (I) populations in the i^{th} region is given by

$$\tau\frac{dE_i}{dt} = -E_i(t) + (S_{E_m} - E_i(t))S_E\Big(c_1E_i(t) - c_2I_i(t) + c_5\sum_j A_{ij}E_j(t-\tau_d^j) + P_i(t)\Big) + \sigma w_i(t), \quad (1)$$

$$\tau\frac{dI_i}{dt} = -I_i(t) + (S_{I_m} - I_i(t))S_I\Big(c_3E_i(t) - c_4I_i(t) + c_6\sum_j A_{ij}I_j(t-\tau_d^j) + Q_i(t)\Big) + \sigma v_i(t), \quad (2)$$

where

$$S_{E,I}(x) = \frac{1}{1+e^{(-a_{E,I}(x-\theta_{E,I})}} - \frac{1}{1+e^{a_{E,I}\theta_{E,I}}}, \quad (3)$$

c_5 and c_6 are excitatory and inhibitory coupling parameters, respectively, which we optimize for a given individual, τ_d represents the distance-based time delay between regions, $\tau = 8$ ms is a time constant, and w_i and v_i are additive noise. The elements of the coupling matrix A describe the connectivity between regions i and j derived from images of an individual's brain, and $P_i(t)$ and $Q_i(t)$ represent the external inputs to excitatory and inhibitory states, respectively. Other constants in the model are biologically derived as described in [1, 2].

References

[1] S. F. Muldoon, F. Pasqualetti, S. Gu, M. Cieslak, S. T. Grafton, J. M. Vettel, and D. S. Bassett. Stimulation-based control of dynamic brain networks. *PLoS Comput Biol*, 12(9):e1005076, 2016.

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

A SEMIDEFINITE PROGRAM FOR STRUCTURED BLOCKMODELS

David Choi

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

Semidefinite programs (SDPs) have recently been developed for the problem of community detection, which may be viewed as a special case of the stochastic blockmodel. Here, we develop a semidefinite program that can be tailored to other instances of the blockmodel, such as non-assortative networks and overlapping communities, and can also approximate latent space models.

For this SDP, we give theorems establishing label recovery in sparse settings, with conditions that are analogous to recent well-known results for community detection. In misspecified settings where the data is not generated by a blockmodel, we give an oracle inequality that bounds excess risk relative to the best blockmodel approximation.

When the assumed blockmodel exhibits symmetry or label-switching ambiguity, the computation time can be significantly reduced by "parameterizing out" the non-identifiable subspace, using a concept known in combinatorics as an association scheme. Simulations and comparison to existing methods are presented for community detection, for overlapping communities, and for latent space models.

A preprint can be found on Arxiv [1].

Idea of the Semidefinite Program

The main idea is the following. In a K-class stochastic blockmodel, we can encode the latent classes by an indicator matrix $Z \in \{0,1\}^{n\times K}$, given by

$$Z_{ik} = \begin{cases} 1 & \text{if node } i \in k\text{th class} \\ 0 & \text{otherwise,} \end{cases},$$

so that the matrix $ZZ^T \in \{0,1\}^{n\times n}$ satisfies

$$[ZZ^T]_{ij} = \begin{cases} 1 & \text{if nodes } i \text{ and } j \text{ are in the same class} \\ 0 & \text{otherwise.} \end{cases}$$

This matrix encodes whether two nodes are in the same class (but not which class they are in).

Previous semidefinite programs worked with a relaxation of the matrix ZZ^T. In this work, we consider a different approach, in which we relax the matrix $\operatorname{vec}(Z)\operatorname{vec}(Z)^T$ instead. Unlike ZZ^T, this matrix is able to fully encode the latent class assignments. As a result, we will show that our approach can be used for any blockmodel.

Theorems for Estimation

We will show that for blockmodels where the average degree is bounded above some constant, the semidefinite program recovers the labels, with missclassification rate bounded by $O\left(1/\sqrt{\text{avg. degree}}\right)$. This implies "weak convergence" for bounded degree graphs, and a vanishing missclassification rate as the degree $\to \infty$.

If the data is not generated by a blockmodel, but instead $A_{ij} \sim \text{Bernoulli}(P_{ij})$ for some arbitrary matrix $P \in [0,1]^{n\times n}$, we will show that the SDP can be used to "denoise" the adjacency matrix and estimate P, with estimation error converging to the best blockmodel approximation to P. Among other things, this suggests that the SDP can be used to approximate latent space models by "discretizing" the underlying latent space.

Optimization

The matrix $\operatorname{vec}(Z)\operatorname{vec}(Z)^T$ is much larger than ZZ^T, having nK rows and columns. At first glance, this suggests that the SDP will be very slow to solve compared to previous versions. However, the semidefinite program will often be highly structured. In particular, whenever the stochastic blockmodel exhibits any type of "label-switching" ambiguity, we will show the non-identifiable subspace can be removed, resulting in a lower dimensional optimization problem. As a result, for many blockmodel types the new SDP will be roughly the same order complexity as previous SDP approaches. In simulation, solving the SDP was practically "fast" for networks with 1000-1500 nodes.

References

[1] D. Choi. A semidefinite program for structured blockmodels. *arXiv preprint arXiv:1611.05407*, 2016.

PARTITIONS, CLUSTERING, AND COMMUNITIES IN MULTIPLEX NETWORKS

Daryl R. DeFord and Scott D. Pauls

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

In an explosion of research over the last decade, numerous authors have adapted many single-layer analytic techniques to the context of multi-layer networks. Detecting and illuminating meso-scale structure within networks — whether identified as communities, partitions, or otherwise — using clustering techniques is a central technique in network analysis as it provides a path to recognize the most salient features of network structure and associated dynamics. In the context of multiplex networks — ones with a single node set but multiple, functionally distinct layers of edges — clustering techniques have proved difficult to extend generally. Work in several directions has been successful in extending ideas in community detection to multi-layer networks but has also revealed a basic stumbling block: many single-layer clustering techniques rely on the analysis and manipulation of structural models of the network and there are a number of possible choices for structural representations of multi-layer networks.

The goal of our work is to explore the impact of different structural representations of multiplex networks in the context of a popular clustering methods, spectral clustering. We focus on two structural models, the diagonal supra-adjacency representation [3] and a model motivated by a multiplex dynamical model [2]. The two models take different approaches to linking the layers into a coherent whole. The first models layer interactions by explicitly including structural ties between copies of the same node in the different layers. The second models these interactions by allowing copies of distinct nodes on different layers to directly interact with one another.

To test these different structural representations, we create several families of synthetic networks on which we use our detection algorithm. First, we create multiplex networks with Erdős-Rényi layers to test cases where we expect no community structure and hence node copies should group together. Second, we create identical planted communities on the layers using a stochastic block model to test a case where any reasonable method should detect the communities. Third, we create different planted communities on the layers to test cases where the overall community structure might be ambiguous, reflecting modeling choices. The last two families are similar to the generative models of meso-scale structure recently put forward in [1].

We find that our two structural models perform differently under these tests. For the diagonal supra-adjacency model, the weight of the inter-layer connections plays a significant role in its effectiveness, which follows from work in Ref. [4]. In spectral clustering, small weight values lead to the layers migrating to different clusters while the method generally leads to the expected results for larger weights. Results for the third experiment are consistent with spectral estimates which show that for large weights the spectrum is linked to the spectrum of the aggregation of the layer networks [5]. The second model, in contrast, performs well on our tests across all parameters. For the last case, the clusterings depend on modeling choices more closely, finding the planted communities on one of the layers based on the preferences encoded in the model parameters.

Our results point to significant consequences of using different structural models, indicating that unlike the single-layer case where a single structural model is appropriate for the majority of applications, multiplex models require more careful consideration.

References

[1] M. Bazzi, L. G. S. Jeub, A. Arenas, S. D. Howison, and M. A. Porter. Generative Benchmark Models for Mesoscale Structure in Multilayer Networks. *ArXiv e-prints*, 2016. 1608.06196.

[2] D. R. DeFord and S. D. Pauls. A new framework for dynamical models on multiplex networks. *arXiv eprints*, 2015. 1507.00695.

[3] S. Gómez, A. Díaz-Guilera, J. Gómez-Gardeñes, C. J. Pérez-Vicente, Y. Moreno, and A. Arenas. Diffusion Dynamics on Multiplex Networks. *Physical Review Letters*, 110(2):028701, Jan. 2013.

[4] F. Radicchi and A. Arenas. Abrupt transition in the structural formation of interconnected networks. *Nature Physics*, 9(11):717–720, Sept. 2013.

[5] A. Solé-Ribalta, M. De Domenico, N. E. Kouvaris, A. Díaz-Guilera, S. Gómez, and A. Arenas. Spectral properties of the laplacian of multiplex networks. *Phys. Rev. E*, 88:032807, Sep 2013.

GRAPH MATCHING THE MATCHABLE NODES WHEN SOME NODES ARE UNMATCHABLE

Vince Lyzinski, Daniel L. Sussman

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Graph matching is the problem of finding a latent correspondence between the nodes in two graphs with the same node set. In many realistic applications, only a *core* subset of the nodes are matchable in each graph, with the remaining nodes being *junk*, only participating in one of the two graphs. Under a statistical model for this situation, we show that correctly matching the core nodes is still possible, provided the number of junk nodes does not grow too rapidly [3].

Abstract

When studying more than one network, it is often fruitful to exploit any underlying correspondence between the nodes in the network, increasing the suite of available tools for analysis. When this correspondence is latent, one can attempt to estimate the correspondence by solving the graph matching problem [1]: For two adjacency matrices $A, B \in \{0,1\}^{n\times n}$, the graph matching problem is to find

$$\text{argmin}_{P\in\mathcal{P}} \|A - PBP^T\|_F^2,$$

where $\mathcal{P}$ is the set of permutation matrices. Assuming the true latent correspondence is the identity mapping, we can statistically model the two graphs by introducing positive correlations between A_{ij} and B_{ij}. In this situation it has been shown that provided the correlation and sparsity do not decay too rapidly, the latent correspondence can be correctly estimated for large graphs [2], computational challenges not withstanding.

As presented above, we assume that all nodes in A have a match in B, and vice versa, but realistically we expect that only some nodes will have a match—the *core* nodes $\mathcal{C} \subset [n]$—and some nodes will only participate in one network—the *junk* nodes $\mathcal{J} = [n] \setminus \mathcal{C}$. We model this situation by imposing that $\text{corr}(A_{ij}, B_{ij}) = 0$ unless both $i, j \in \mathcal{C}$. Consider that for all $i, j \in [n]$, with $i < j$, $A_{i,j}, B_{ij} \sim \text{Bern}(\Lambda_{ij})$ independently across i, j and with $\text{corr}(A_{ij}, B_{ij}) = R_{ij}$.

In the challenging yet simple case where all $\Lambda_{ij} = \lambda$ and $R_{ij} = \rho\mathbf{I}\{i, j \in \mathcal{C}\}$, we are able to show that if the number of junk nodes $n_J = |\mathcal{J}|$ is sublinear in the number of core nodes $n_C = |\mathcal{C}|$, the minimizing permutation P will correctly match core nodes, with $P_{ii} = 1$ for all $i \in \mathcal{C}$. We can also allow for more complex situations, where both the edge probabilities Λ_{ij} and the edge-wise correlations R_{ij} are allowed to vary, but the resulting bounds decay to $|\mathcal{J}| \leq \sqrt{|\mathcal{C}|}$. In the situation where A_{ij} and B_{ij} are not assumed to be identically distributed, our proof technique and methods are less powerful. This is due the possibility that with enough junk nodes, some junk nodes in the first graph can *behave* more like core nodes than junk node in the second network.

In our presentation, we will demonstrate these results in simulation and we will also investigate matching graphs derived from Twitter in two separate months. We synthetically create the core-junk situation for the Twitter data by keeping a subset of the nodes in both graphs and the remaining nodes are each kept in only one graph. While solving the graph matching problem exactly is computationally intractable, we use a state-of-the-art approximate algorithm which exploits a relaxation of the constraints to make the problem continuous [4]. We also use seeds, a set of nodes where the correspondence is known, to improve the accuracy and computation time for our simulations.

After matching, we also investigate ways of clustering nodes into core and junk sets and ranking which nodes are mostly likely to be core nodes. Extensions beyond the case that A and B are the same size are also being explored by appropriately padding the smaller matrix.

References

[1] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.

[2] V. Lyzinski, D. E. Fishkind, and C. E. Priebe. Seeded graph matching for correlated Erdos-Renyi graphs. *Journal of Machine Learning Research*, 15:3513–3540, 2014.

[3] V. Lyzinski and D. L. Sussman. Graph matching the matchable nodes when some nodes are unmatchable. 5 May 2017.

[4] J. T. Vogelstein, J. M. Conroy, V. Lyzinski, L. J. Podrazik, S. G. Kratzer, E. T. Harley, D. E. Fishkind, R. J. Vogelstein, and C. E. Priebe. Fast Approximate Quadratic Programming for Graph Matching. *PLoS ONE*, 10(04), 2014.

RIGIDITY PERCOLATION IN COMPOSITE MATERIALS

Samuel Heroy, Dane Taylor, F. Bill Shi, Peter J. Mucha, & M. Gregory Forest

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Mechanical percolation is a phenomenon in materials processing wherein 'filler' rod-like particles are incorporated into polymeric materials to enhance the composite's mechanical properties. Experiments have well characterized a nonlinear phase transition from floppy to rigid behavior at a threshold filler concentration, but the underlying mechanism is not well understood. We develop and utilize an iterative graph compression algorithm to demonstrate that this experimental phenomenon coincides with the formation of a spatially extending set of mutually rigid rods ('rigidity percolation').

Background

Nanoscopic to microscopic particles of high aspect ratio are routinely incorporated into polymeric host materials to enhance attributes such as electrical or thermal conductivity, charge storage, and mechanical properties [1, 2, 3, 4, 5]. Experimentally, such composite materials typically exhibit a nonlinear response with respect to the density of rods or filaments: the property gain scales linearly with rod density at low densities, then soars as rod density approaches and exceeds a critical threshold. For conductivity, this sharp transition with conducting rods in a poorly conducting polymer is understood as a contact percolation phase transition within the rod network. A sharp rise in mechanical stability, however, occurs at volume fractions well beyond the contact percolation threshold. This *mechanical percolation* phenomenon has been experimentally characterized in many rod composites, but the underlying physical mechanism remains a subject of interest, motivating different approaches including mean-field micromechanics models [1], effective medium theory [2], and for nanoscale rods, incorporation of the interfacial domain between the rods and host polymer [6]. Here, we explore the underlying network structures within the rod phase that generalize contact percolation to *rigidity percolation*, one potential source of mechanical percolation.

Methodology

We model the rod phase in composites as a spatial dispersion of thin rectangles (in two dimensions) or cylinders (in three dimensions). For tractability, we model each rod contact point as a *hinge*, assuming for simplicity that friction and other attractive forces keep the rods in contact, while allowing them to rotate freely about their point of contact. It is our hypothesis that mechanical percolation in nanorod composites may be successfully modeled and analyzed through the formation of a spatially-extended set of connected rods that are mutually rigid, which we refer to as a *spanning rigid cluster*.

Rigid Graph Compression (RGC)

Our rigidity analysis is based on the identification of certain topological motifs—i.e. three rods intersecting pairwise form a single rigid body—which we prove analytically using *rigidity matroid theory* [7]. These rigidity motifs apply hierarchically, and so we integrate them into an iterative algorithm—Rigid Graph Compression (RGC)—which we use to decompose large ensembles into mutually rigid sets of rods (see Figure 1).

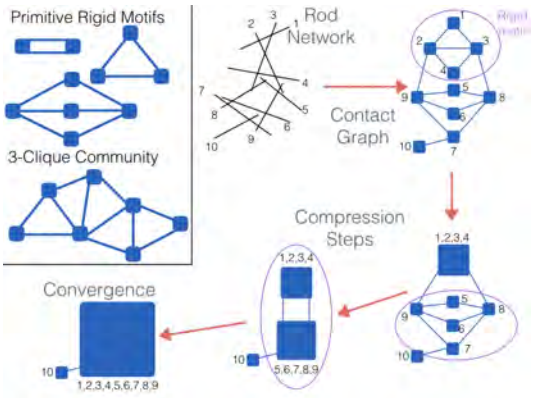


Figure 1: RGC is used to identify a 9-rod rigid cluster in a 2D rod dispersion. First, this dispersion is transformed into a graph wherein rods are represented as nodes, which share an edge if they intersect (or nearly intersect in 3D). Then, three primitive 'rigidity motifs'—provably rigid subgraphs (see inset)—are used to iteratively identify and compress rigid bodies, which may be rods or already identified mutually rigid sets of rods. The 3-clique community (inset) is a non-primitive motif of computational value, which can constructed using the 2- and 3-body motifs.

We first verify our method in 2D, finding that three rigidity motifs are sufficient to estimate the rigidity percolation threshold and correlation length parameter associated with the formation of a spanning rigid cluster. Then, we apply RGC to 3D rod networks and demonstrate the formation of a spanning rigid cluster using four motifs. As no other rigidity detection analysis has been applied to 3D rod networks, we will (in the future) verify the efficacy of this approach by comparing to dynamical simulations of the composite system.

References

[1] S. Baxter and C. Robinson. Pseudo-percolation: Critical volume fractions and mechanical percolation in polymer nanocomposites. *Compos. Sci. Technol.*, 71(10):1273–1279, 2011.

[2] M. Niklaus. Electrical conductivity and young's modulus of flexible nanocomposites made by metal-ion implantation of polydimethylsiloxane: The relationship between nanostructure and macroscopic properties. *Acta Mater.*, 59(2):830–840, 2011.

[3] F. Shi, S. Wang, M. Forest, and P. Mucha. Percolation-induced exponential scaling in the large current tails of random resistor networks. *Multiscale Model. Simul.*, 11(4):1298–1310, 2013.

[4] F. Shi, S. Wang, M. Forest, P. Mucha, and R. Zhou. Network-based assessments of percolation-induced current distributions in sheared rod macromolecular dispersions. *Multiscale Model. Simul.*, 12(1):249–264, 2014.

[5] X. Zheng, M. Forest, R. Vaia, M. Arlen, and R. Zhou. A strategy for dimensional percolation in sheared nanorod dispersions. *Adv. Mater.*, 19(22):4038–4043, 2007.

[6] R. Qiao and L. Brinson. Simulation of interphase percolation and gradients in polymer nanocomposites. *Compos. Sci. Technol.*, 69(3-4):491–499, 2009.

[7] B. Hendrickson. Conditions for unique graph realizations. *SIAM J. Comput.*, 21(1):65–84, 1992.

MEASURING AND MONITORING COLLECTIVE ATTENTION DURING DISASTERS

Yu-Ru Lin, Xingsheng He

SIAM Workshop on Network Science 2017
July 13-14 · Pittsburgh

Summary

We propose an *attention shift network* framework to systematically analyze the dynamics of collective attention in response to real-world exogenous shocks such as disasters. Through tracing hashtags that appeared in Twitter users' complete timeline around several violent terrorist attacks in 2015 and 2016, we study the properties of network structures and reveal the temporal dynamics of the collective attention across multiple disasters. Further, to achieve a more efficient monitoring of the collective attention dynamics, we propose an effective stochastic graph sampling approach that accounts for the users' hashtag adoption diversity and data variability.

Extended Abstract

In recent years, there has been growing interest in the use of social media in crisis response, with scope ranging across natural disasters, terrorist attacks, and political riots. During these events, the flow and flood of information can easily lead to a poverty of attention and thus creates a need to allocate such attention efficiently for affected communities. Consequently, a systematic understanding of attention dynamics at the collective level within disaster context serves as the basis for scheduling effective crisis communications, facilitating timely crisis response such as just-in-time warning and evacuation.

In this work, we seek to quantitatively capture the collective attention shift under exogenous shocks, specifically disaster events, by using Twitter users' communication streams. Fig. 1 illustrates the collective attention *before* and *after* the 2015 Paris attacks event based on how Paris users shift their attention to various topics – captured by the use of different *hashtags*. Before the event, users' attended topics exhibited a salient community structure, reflecting their scattered attention among various topics. After the event happened, a few hashtags became the hubs that suddenly appeared in many users' tweets. Such sudden change in users' attended topics at the collective level is referred to as "collective attention shift."

We introduce a new framework for capturing collective attention shift. We illustrate our framework by using a large corpus of twitter communications centered around multiple shocking terrorist attacks in 2015 and 2016. We employ hashtags as a proxy for users' attended topics, and utilize the hashtag adoption sequence from a user's tweet timeline as the trace of his/her attention shift process. We construct an *attention shift network* or *attention graph* to represent the attention shift process at the collective level.

(a) Pre-event (11/12/2015) (b) Post-event (11/13/2015)

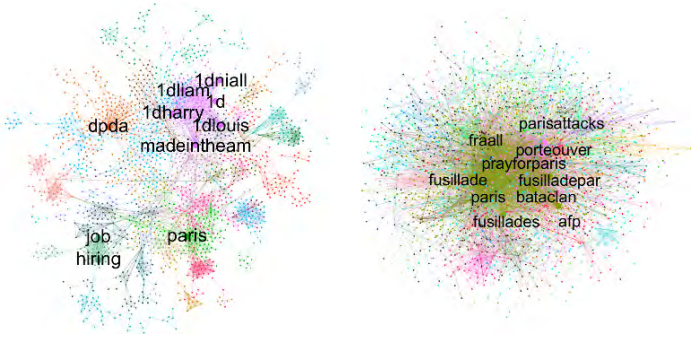


Figure 1: Collective attention *before* and *after* the 2015 Paris attacks event captured by attention shift networks.

Based on this network representation, we quantify the structural change of collective attention shift and further examine data sampling schemes that can capture the structural change in a cost-effective manner. Our study of the collective attention during multiple shocking terrorist attack events in 2015 and 2016 and reveals several properties of network structures and temporal dynamics that are consistent across events.

We formulate a new problem for efficient monitoring of the collective attention dynamics, and we propose a cost-efficient sampling strategy that takes the users' hashtag adoption frequency, connectedness and diversity into account, with a stochastic sampling algorithm to cope with the variability of the sampling targets. We show that our proposed sampling approach outperforms several alternative methods in both retaining the network structures and preserving the information with a small set of sampling targets, suggesting the utility of our method in various realistic settings.

INVITED TALK 2: STEFANO ALLESINA, UNIVERSITY OF CHICAGO

Higher-order interactions stabilize dynamics in competitive networks

Networks have a long history in ecology, and have been used to represent consumer-resource (food webs), mutualistic (pollination, seed-dispersal),and competitive interactions. In these ecological networks, species are nodes and edges represent interactions between pairs of species. Early on, ecologists realized that pairwise interactions might not be sufficient to describe the intricacies of ecological systems, but so far empirical evidence for higher-order interactions has been scant. Moreover, we lack an understanding of how these higher-order interactions would affect population dynamics. I present a simple model of competition in which higher-order interactions dramatically change dynamics: when species interact in pairs, instability prevents the persistence of large communities; when species can interact in triplets, quadruplets, etc., Dynamics are stable and we can build persisting communities containing an arbitrary number of species.

DETANGLING THE HAIRBALL: LESSONS FROM THE DREAM 2016 DISEASE MODULE DETECTION CHALLENGE

Lenore J. Cowen

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

In Fall of 2016, the DREAM Disease Module Identification challenge [5] invited participants to apply community detection or graph clustering methods to predict functional modules in heterogeneous anonymized genomic networks based solely on network structure. Predictions were scored based on the enrichment of modules for genes that had been implicated in human disease based on GWAS studies. We produced the winning entry [3] from a field of 42 teams by combining our novel DSD metric for "detangling" the network [1, 2] with standard off the shelf methods for spectral clustering. We discuss the dataset (which itself may be of broad interest to this community), how the challenge was set up, what it was like to participate, and what challenges and questions still remain for uncovering disease-related communities in biological networks.

Introduction

A great deal of high-throughput information about human genes can be represented in the form of gene-gene or protein-protein association networks. It is well known that such networks have a high degree of modularity, and that the corresponding modules often comprise genes or proteins that are involved in the same biological function. The DREAM challenge was designed as a community effort to extensively benchmark different methods and parameter settings to best reveal biologically relevant modules across diverse types of genomic networks. The contest dataset comprised six different anonymized networks derived from human genomic data, summarized in Table 1.

Results were evaluated based on the number of discovered modules that were statistically significantly associated with complex traits and diseases. To this end, the challenge organizers collected over 200 GWAS datasets that were associated with a broad range of complex traits and diseases, half provided for parameter tuning in training rounds; the rest for evaluation of the submissions.

Network	→	# Nodes	# Edges	Edge weight
PPI-1	N	17,397	2,232,405	Confidence
PPI-2	N	12,420	397,309	Confidence
Signaling	Y	5,254	21,826	Confidence
Expression	N	14,679	1,000,000	Correlation
Cancer	N	14,678	1,000,000	Correlation
Homology	N	10,405	4,223,606	Confidence

Results

Our team's approach to the challenge was based on our DSD spectral graph metric of Cao et al. [1, 2]. We preprocessed all networks using DSD to "detangle" the networks, and then applied off the shelf clustering and community detection methods to the detangled networks. [4, 6]. This approach won on the subchallenge that considered each of the six network separately; and performed in the top cohort on the subchallenge that integrated information across the six heterogeneous networks. We discuss our winning strategy, the interesting data sets, and what seem to be both the strengths and the limitations of current methods for these biological datasets.

.

References

[1] M. Cao, C. M. Pietras, X. Feng, K. J. Doroschak, T. S. ffner, J. Park, H. Zhang, L. J. Cowen, and B. Hescott. New directions for diffusion-based prediction of protein function: incorporating pathways with confidence. *Bioinformatics*, 30:i219–i227, 2014.

[2] M. Cao, H. Zhang, J. Park, N. M. Daniels, M. E. Crovella, L. J. Cowen, and B. Hescott. Going the distance for protein function prediction. *PLOS One*, 8:e76339, 2013.

[3] J. Crawford, J. Lin, X. Hu, B. Hescott, D. Slonim, and L. J. Cowen. A double spectral approach to DREAM11 subchallege 3. https://www.synapse.org/# !Synapse:syn6156761/wiki/407453. Manuscript in preparation. See also: http://dreamchallenges.org/all-stars.

[4] M. Girvan and M. E. Newman. Community structure in social and biological networks. *PNAS*, 99(12):7821–7826, 2002.

[5] D. Marbach et al. Disease module identification DREAM challenge. https://www.synapse.org/# !Synapse:syn6156761/wiki/. Manuscript in preparation.

[6] F. Pedregosa et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830, 2011.

ACCURATE AND EFFICIENT PREDICTIONS OF FIRST-PASSAGE TIMES IN SPARSE DISCRETE FRACTURE NETWORKS USING GRAPH-BASED REDUCTIONS

Jeffrey D. Hyman, Aric Hagberg, Gowri Srinivasan, Jamaludin Mohd-Yusof, Hari Viswanthan

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

We present a graph-based methodology to reduce the computational cost of obtaining first passage times through sparse fracture networks. We derive graph representations of generic three-dimensional discrete fracture networks (DFN) using the DFN topology and flow boundary conditions. Subgraphs corresponding to the k shortest loopless paths between the inflow to outflow boundaries are identified and transport on their equivalent subnetworks is compared to transport through the full network. The number of paths included the subgraphs is based on the scaling behavior of the number of edges in the graph with the number of shortest paths. First passage times through the subnetworks are in good agreement with those obtained in the full network, both for individual realizations and in distribution. Accurate estimates of first passage times are obtained with an order of magnitude reduction of CPU time and mesh size using the proposed method.

We generated 100 three-dimensional generic discrete fracture networks. The fracture networks are fairly sparse but dense enough that there are multiple paths between the inflow and outflow boundaries. Steady state conditions are numerically determined to obtain the fluid velocity field within each network. A graph representation G of each DFN F is constructed using the network topology. We also include source and target vertices into G to incorporate flow direction. The mapping is bijective, so every subgraph $G' \subseteq G$ has a unique pre-image F' in the fracture network,

Subgraphs G', along with their equivalent subnetworks F', corresponding to the union of the edges in k-shortest loopless paths from the source to target are identified in each graph. All edges in G have unit weight, so these paths correspond to the fewest number of edges between the source and the target. The pre-image of this subgraph, its equivalent fracture subnetwork F', is the fewest number of intersections, and thus connected fractures, spanning the inflow and outflow boundaries. We also consider the 2-core of the graph, which is an upper bound on the union of loopless paths from source to target. Figure 1 shows three subnetworks (top) and their subgraphs (middle). Semi-transparent vertices indicate fractures that have been eliminated from the fracture network.

Accuracy of the method is determined by comparing the first passage times of a solute transported through the full network $\hat{\tau}$ and the subnetworks corresponding to the k shortest paths $\hat{\tau}'$. While the shortest path requires the smallest CPU times, it provides the worst estimates of first passage times. Using the ten shortest paths requires slightly more CPU time, but the predictions of first passage times are significantly improved. The primary paths through the network, discussed above, are included in the first ten shortest paths for all networks. The 2-core of the graph, provided the best predictions of first passage times. However, the CPU time required for computation on the 2-core subnetwork was 75% of that needed for the full network, underscoring the trade-off between accuracy and efficiency.

(a) Shortest Path (b) Ten Shortest Paths (c) Network 2-Core

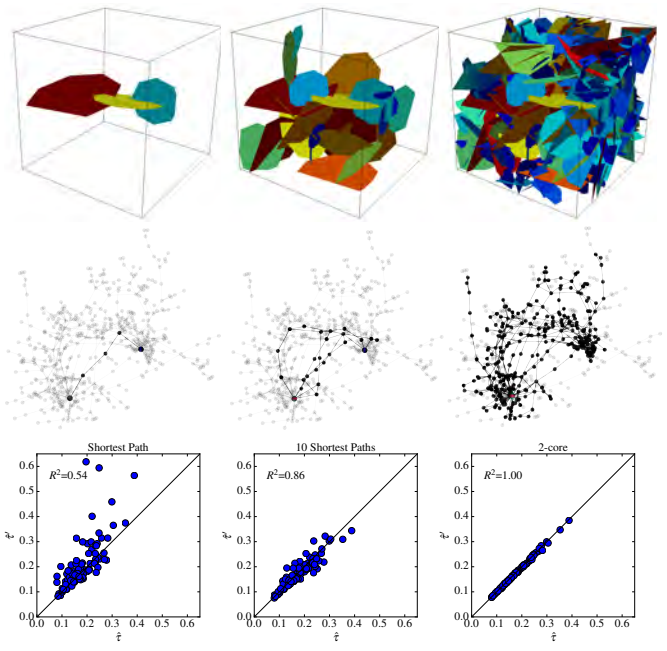


Figure 1: Subnetworks (top) and subgraphs (middle) derived from the full DFN. (a) The shortest path through the network, (b) the union of the ten shortest paths in the network and, (c) the 2-core. Semi-transparent vertices denote fractures that have been eliminated from the fracture network. The first passage-times (bottom) get more accurate as more edges are included.

INTERPLAY BETWEEN SYSTEM RATIONALITY AND TOPOLOGICAL STRUCTURE OF SUPPLY CHAIN NETWORKS

Supun Perera, Dharshana Kasthurirathna, Michael Bell

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Recent developments in network science have enabled researchers to investigate the topological structure of Supply Chain Networks (SCNs) [1]. A typical SCN model consists of nodes which represent individual firms and links which represent various interactions between firms. In this study, we present a topological model of bounded rationality for SCNs.

Methodology

In order to generate an ensemble of network topologies representative of SCNs, we use the *Log Normal Fitness Attachment* (LNFA) mechanism [2]. The tunable σ parameter of the lognormal distribution offers considerable flexibility to this model, as it can be varied to generate a wide spectrum of network topologies. Using the LNFA protocol, by varying σ, we generated $2,000$ networks (each with $1,000$ nodes), with scale-free exponents ranging between 1.5 and 4.5.

Inspired by the work of [4], here we argue that there could be a direct relationship between the amount of social interaction of a particular player and their level of bounded rationality. Accordingly, we modelled the rationality of each firm, as a monotonically increasing linear function of their degree, with a network wide parameter to control the responsiveness of rationality to degree.

Nash Equilibrium (NE) assumes that players behave perfectly rationally. However, in reality, players are only boundedly rational due to limitations in information availability, computational time and cognitive capacity. Quantal Response Equilibrium (QRE) offers a direct way to model games with *noisy* strategies, by using logit probabilistic choice functions. For instance, the logit formation used by [3] is parametrised in a way that is directly analogous to the bounded rationality interpretation. As the rationality parameter is varied from zero to infinity, the choice behaviour of the agents moves from random to fully rational. In establishing the logit QRE for each interaction, we used the Prisoner's Dilemma game.

The Jensen-Shannon divergence (JSD) is generally used to measure the divergence between two probability distributions. Accordingly, we use the JSD between the QRE and NE of each interaction (i.e. link) in the SCN, as an indicator of the rationality of that interaction. Then we averaged the JSD values over all interactions in the SCN to gauge the overall system rationality. This is based on the game theoretic assumption that proximity to NE is an indicator to a certain player's rationality [4].

Results and Discussion

The scatter plot for the average JSD against the scale-free exponent values, of the networks considered, indicates that when the scale-free exponent is below 2.5, the cumulative interactions within each SCN rapidly approach the NE. Since $\gamma = 2$ is the boundary between hub and spoke ($\gamma<2$) and scale-free ($\gamma>2$) network topologies [5], it suggests that when the individual node rationality correlates with its topological degree, it may give rise to a hub and spoke network topology in a competitive strategic decision making environment.

Acknowledgements

This work has been funded by the Australian Research Council (ARC) under grant DP140103643.

References

[1] Stephen P Borgatti and Xun Li. On social network analysis in a supply chain context. *Journal of Supply Chain Management*, 45(2):5–22, 2009.

[2] Shilpa Ghadge, Timothy Killingback, Bala Sundaram, and Duc A Tran. A statistical construction of power-law networks. *International Journal of Parallel, Emergent and Distributed Systems*, 25(3):223–235, 2010.

[3] Richard D McKelvey and Thomas R Palfrey. Quantal response equilibria for extensive form games. *Experimental economics*, 1(1):9–41, 1998.

[4] Dharshana Kasthurirathna and Mahendra Piraveenan. Emergence of scale-free characteristics in socio-ecological systems with bounded rationality. *Scientific reports*, 5, 2015.

[5] Albert-László Barabási. Network science book. *Boston, MA: Center for Complex Network, Northeastern University. Available online at: http://barabasi. com/networksciencebook*, 2014.

SPATIAL AND SOCIAL ORGANIZATION OF AN ANT COLONY: A NETWORK ANALYSIS

Ewan Colman, Andreas Modlmeier, David Hughes, Shweta Bansal

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Ant colonies exhibit a remarkable social dynamic in which each ant specializes in one particular function. The roles that have been observed are often analogous to human activities, there are nurses, workers, foragers, guards and so on. To perform these tasks effectively there must be sufficient communication between ants of different types; the foragers need to react when other ants are in need of food, the guards need to be called to arms when a threat appears, and each ant needs must be aware of what the others are doing in order to decide what she herself should do. By observing interactions and mapping their communication networks we can begin to understand how ants self-organize to achieve balance and efficiency.

Abstract

One way to find out how ant society is structured is to manipulate its environment and observe how the colony adapts. In a laboratory experiment we introduced a colony into an artificial nest box that was barely large enough to contain them, then, after observations had been made, the box was extended to four times its original size. We tracked the locations of all 80 ants and recorded each trophallaxis (food-sharing) interaction over a 4-hour period. At first it seemed like the ants changed their behaviors in response to the changes in nest space; they adapted to the larger nest box by segregating into two distinct spatial regions, prompting the question: how does their spatial organization affect communication and food distribution throughout the colony?

To answer this we constructed two types of network. In the first, ants are nodes and the weight of each edge represents how similar they are in regards to their movement patterns. The second is a temporal network constructed from the trophallaxis interactions. We analysed community structure, path lengths, path durations and communicability. We also developed a mathematical model that considers the heterogeneity in contact rates between different pairs of nodes and fit the model to the trophallaxis data and to several other animal and human contact networks.

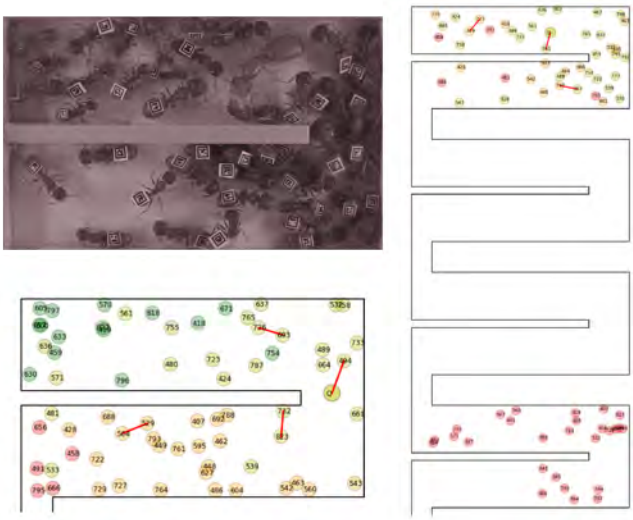


Figure 1: Inside the ant nest. The image in the top left corresponds to the upper nest chamber in the large nest. Images are not drawn to scale. The colors represent which group it was assigned by the process described. The red lines represent a trophallaxis interaction.

We conclude that ants exhibit a remarkable consistency in their social structure even when tested by extreme changes to their environment, in particular:

- The spatial and social structure of the ant colony is robust against changes to their nest size.
- Spatial structure and social structure, as observed in the trophallaxis network, are co-dependent.
- The rate at which food and information spreads through the network of trophallaxis interactions is not affected by the size and density of the nest.
- Ant populations are more homogeneously mixed than human and other animal populations. This aspect of their behavior is not affected by the nest size.

NONBACKTRACKING WALK CENTRALITY FOR DIRECTED NETWORKS

Francesca Arrigo, Peter Grindrod, Desmond J. Higham, and Vanni Noferini

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We argue that nonbacktracking walks are relevant and useful for centrality computations. By exploiting an analytic expression for the appropriate generating function, we show that Katz centrality can be made nonbacktracking on directed graphs at no extra cost. The range of values available for the downweighting parameter is found to be determined by the spectrum of a three-by-three block matrix involving the original adjacency matrix. Theoretical and numerical examples will be used to the illustrate the benefits of not tracking back.

Background and Problem Setting

Let $G = (V, E)$ be a loopless, unweighted digraph with n nodes and no multiple edges. Let $A \in \mathbb{R}^{n\times n}$ be its adjacency matrix, $D \in \mathbb{R}^{n\times n}$ be the diagonal matrix whose diagonal entries are $D_{ii} = (A^2)_{ii}$, $S \in \mathbb{R}^{n\times n}$ be the matrix whose entries are defined as $S_{ij} = A_{ij}A_{ji}$, and I be the identity matrix. A walk, i.e., traversal through a sequence of (possibly repeated) nodes, is said to be *backtracking* if it contains at least one node subsequence of the form $i\,\ell\,i$, and *nonbacktracking* (NBTW) otherwise. NBTWs have typically been studied on undirected networks [1, 3, 4], but the definition continues to make sense in the directed case [5]. It is intuitively reasonable to consider a version of Katz [2] based on NBTWs. Let us denote by $p_r(A)$ the matrix whose (i, j)th entry counts the number of NBTWs of length r from node i to node j.

We define and study the NBTW centrality measure $\mathbf{b} = \left(\sum_{r=0}^{\infty} t^r p_r(A)\right)\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector of all ones and $t > 0$. Here $b_i > 0$ gives the centrality of node i.

NBTW-based Centrality

Theorem 1 ([5]) *Let $\phi(A, t) := \sum_{r=0}^{\infty} t^r p_r(A)$ and let $M(t) = I - At + (D - I)t^2 + (A - S)t^3$. Then, within the radius of convergence, $M(t)\phi(A, t) = (1 - t^2)I$.*

Theorem 1 shows that the NBTW centrality measure $\mathbf{b}$ satisfies $M(t)\mathbf{b} = (1-t^2)\mathbf{1}$. Since $M(t)$ has the same sparsity as $I - tA$, we see that NBTW centrality for digraphs may be computed at least as cheaply as Katz [2].

Radius of Convergence

Theorem 2 *The power series $\sum_{r=0}^{\infty} t^r p_r(A)$ converges if $0 < t < 1/\rho(C)$, where $\rho(C)$ is the spectral radius of*

$$C := \begin{bmatrix} A & (I - D) & (S - A) \\ I & 0 & 0 \\ 0 & I & 0 \end{bmatrix} \in \mathbb{R}^{3n\times 3n}.$$

Theorem 2 determines a suitable range for the parameter t. The restriction is less severe than the Katz version, $0 < t < 1/\rho(A)$, and the difference can be dramatic.

We will also show that removing certain types of nodes from the network does not affect the spectral radius of C; thus, the cost of computing $\rho(C)$ can be significantly reduced. Moreover, the same pruning operations can be used to speed up the linear system solve.

By considering the limit $t \to 1/\rho(C)$ from below, we also obtain a generalization to the directed case of the nonbacktracking eigenvector centrality measure from [3].

We will give theoretical examples to show that the new NBTW centrality measure (a) can avoid unwanted *localizaton* effects present in Katz, and (b) performs in an intuitively more reasonable manner than its eigenvector counterpart on star-like graphs. Comparisons will also be given for real networks.

Acknowledgements

The work of FA and DJH was supported by the EPSRC under grant EP/M00158X/1. DJH was also supported by a Royal Society/Wolfson Research Merit Award.

References

[1] P. Grindrod, D. J. Higham, and V. Noferini. The deformed graph Laplacian and its applications to network centrality analysis. *Preprint, submitted*, 2017.

[2] L. Katz. A new index derived from sociometric data analysis. *Psychometrika*, 18:39–43, 1953.

[3] T. Martin, X. Zhang, and M. E. J. Newman. Localization and centrality in networks. *Phys. Rev. E*, 90:052808, 2014.

[4] H. Stark and A. Terras. Zeta functions of finite graphs and coverings. *Advances in Mathematics*, 121(1):124–165, 1996.

[5] A. Tarfulea and R. Perlis. An ihara formula for partially directed graphs. *Linear Algebra and its Applications*, 431:73–85, 2009.

A NEW ALGORITHM MODEL FOR MASSIVE-SCALE STREAMING GRAPH ANALYSIS

Chunxing Yin, Jason Riedy, and David A. Bader

SIAM Workshop on Network Science 2016
July 15-16 · Boston

Summary

Applications in computer network security, social media analysis, and other areas rely on analyzing a changing environment. The data is rich in relationships and lends itself to graph analysis. Traditional static graph analysis cannot keep pace with network security applications analyzing nearly one million events per second[3] and social networks like Facebook collecting 500 thousand comments per second[5]. Streaming frameworks like STINGER[2] support ingesting up three million of edge changes per second but there are few streaming analysis kernels that keep up with these rates. Here we present a new algorithm model for applying complex metrics to a changing graph. In this model, many more algorithms can be applied without having to stop the world.

Non-Stop Streaming Data Analysis Model

In our *non-stop streaming data analysis model*, the input stream keeps making changes to a graph concurrent with analysis algorithms. The algorithms do not have access to the changes explicitly and may or may not encounter changes during their execution. We consider an algorithm valid for our model if it produces a correct result on a graph consisting of the starting graph and *some unspecified subset of concurrent changes*. Clearly, not all algorithms will remain valid in our streaming model, but a surprisingly useful subset are valid. So far, we have shown[4] the following algorithms valid under reasonable assumptions: breadth-first search, triangle counting (modified algorithm), simplified Shiloach-Vishkin connected components, and PageRank.

Algorithms that are not valid often make a decision twice on data that has changed. For example, some algorithms treat high-degree and low-degree vertices differently. If the classification of high-degree or low-degree is not saved, the graph could change to push a vertex from one category to another, and an algorithm could completely miss a vertex. Another example is S. Kahan's connected components algorithm[1]. That algorithm computes the connected components of a reduced graph and then returns to label the original graph. If previously disconnected components are now connected, the labeling procedure could overwrite its own results unpredictably.

In [4], we prove that validity is the strongest form of correctness in our model for algorithms that produce subgraphs (*e.g.* tree building, community extraction) subject to some reasonable assumptions on execution. Fully general analysis would require keeping snapshots or versioned data structures. Neither are feasible at these scales, tens to hundreds of billions of edges, or rates of change, many millions of updates per second.

Extension to Streaming Kernels

Our model can be extended to support streaming updates to analysis results. We assume that *all* changes made during a kernel's execution are recorded and set aside. After execution, a kernel can use those changes to compute focused updates. If concurrent changes again are saved, and if the updating algorithm is valid in our model, the process can repeat to update graph analyses efficiently.

Consider triangle counting. To update the triangle counts, an algorithm can re-compute the triangle count starting from only the vertices changed during its previous execution. This produces a result valid for the "starting graph" by incorporating the prior changes and the unknown subset of concurrent changes.

References

[1] J. W. Berry, B. Hendrickson, S. Kahan, and P. Konecny. Software and algorithms for graph queries on multithreaded architectures. In *2007 IEEE International Parallel and Distributed Processing Symposium*, pages 1–14. IEEE, 2007.

[2] D. Ediger, R. McColl, J. Riedy, and D. A. Bader. STINGER: High performance data structure for streaming graphs. In *The IEEE High Performance Extreme Computing Conference (HPEC)*, Waltham, MA, Sept. 2012. Best paper award.

[3] M. Vallentin, V. Paxson, and R. Sommer. Vast: A unified platform for interactive network forensics. In *13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16)*, pages 345–362, Santa Clara, CA, 2016. USENIX Association.

[4] C. Yin, J. Riedy, and D. A. Bader. Validity of graph algorithms on streaming data. 2017. (in submission).

[5] Zephoria. https://zephoria.com/top-15-valuable-facebook-statistics/ retrieved in January 2017.

ADVERSARIAL ANALYSIS OF COMMUNITY DETECTION

W. Philip Kegelmeyer, Jeremy Wendt, Ali Pinar, Kristen Altenburger

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Community detection is often used to understand the nature of a network. However, there may exist an adversarial member of the network who wishes to evade that understanding. We analyze such a situation, quantifying the efficacy of certain attacks against community detection and providing preliminary results on possible defenses.

A Sample Adversarial Model

Consider a network in which each node has been assigned a starting "temperature"; "hot", "cold", or "unknown", with corresponding values of 1, -1, and 0. The temperature of a community is then the average of its nodes' temperatures. The adversary's goal is to avoid being associated with other hot nodes. The only "attack" permitted, in the current model, is to make links from itself to other nodes, as illustrated in Figure 1, where the green-ringed adversary node has added orange false edges to pull itself into a cooler community.

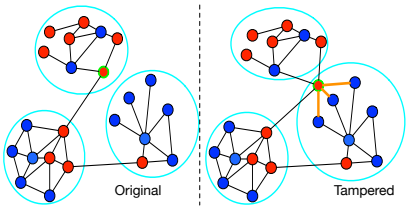


Figure 1: An Example of Temperature Tampering

The adversary's goal for a given network is to reduce its community temperature as much as possible with the fewest added edges. We assume an adversary who has full knowledge of the network and further knows that Louvain will be used as the community detection method.

Attack Evaluation

We abstract a specific "attack" as a list of the network's nodes in the order that the adversarial node will attach to them. We have invented a handful of heuristics to guide the creation of these attacks. We assess each attack by adding N false edges in the indicated order and evaluating the temperature T of the adversary's resulting community. We plot N vs T in "attack efficacy" curves; an example is in Figure 2.

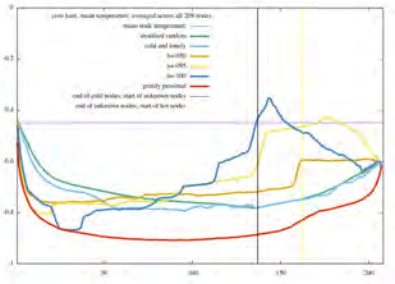


Figure 2: Example Attack Efficacy Curves

How To Defend?

If one could exactly detect the falsely inserted edges, they could be removed, thereby returning the adversary to their "natural" community. We have conducted a preliminary investigation around building machine learning models to that end. We extract various features of each edge (such as the betweenness centrality of the edge's nodes), generate training data with the inserted edges labeled, and use decision tree ensembles to detect them.

As one example result on real data, the original temperature of the adversary was 0.15. After adding 20 false edges, the adversary node was able to lower its temperature to -0.79. Applying an initial machine learning model raised the adversary's temperature back up to -0.02, and so was able to remediate much of the attack. Though to do so, it removed 65 of the network's edges, many more than the 20 false edges, and so altered some of the native structure of the network. Accordingly, current work is addressing both improving the defense models and evaluating the consequences of false alarms in the model.

Supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energys National Nuclear Security Administration under contract DE-AC04-94AL85000.

IDENTIFYING THE COUPLING STRUCTURE IN COMPLEX SYSTEMS THROUGH THE OPTIMAL CAUSATION ENTROPY PRINCIPLE, INFORMATION FLOW AND INFORMATION FRAGILITY

Erik Bollt and Jie Sun

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

The inverse problem of finding networks from data, sometimes called network tomography, has been a long standing and important issue for incorporating the theory of complex systems toward real world applications. Here we take an information theoretic perspective that information flow relates to network structure and dynamics on networks, but it is crucial to have a way to distinguish direct versus indirect influences. We develop Causation Entropy (CSE) to this purpose, and utilize it in a constructive algorithm that we call oCSE for optimal causation entropy.

Description

Inferring the coupling structure of complex systems from time series data in general by means of statistical and information-theoretic techniques is a challenging problem in applied science. The reliability of statistical inferences requires the construction of suitable information-theoretic measures that take into account both direct and indirect influences, manifest in the form of information flows, between the components within the system. In this work, we present an application of the optimal causation entropy (oCSE) principle [2, 3, 5, 4, 1], to identify the coupling structure and jointly apply the aggregative discovery and progressive removal algorithms based on the oCSE principle to infer the coupling structure of the system from the measured data. We will include discussion of examples such as the functional brain network as inferred by fMRI functional magnetic imaging.

Identifying connections in a complex process manifest as causal direct information flowsuggests a new way of detecting and understanding fundamental changes in the dynamical process of a complex system. The question of fragility and robustness concerns how the macroscopic behavior of a system will change in response to local perturbations.We interpret the phrases robust and fragile as a global descriptor of the system, in terms of the change of the information carrying capacity of paths between states of a complex system, due to the loss of a state, or connection, with a corresponding descriptor in terms of information betweenness. Stated more broadly about the interdependencies of complex systems, consider a large-scale process in which minor changes frequently occur, and the question is, can we define and, hence detect, those changes which would render the system effectively different and likewise significantly alter the system performance, before the system might fail. Thus, here we suggest a fragility-robustness duality to detect a tipping point whereby even a minimal detail change can cause a catastrophic systemic outcome.

Acknowledgments

We thank Dr. Samuel Stanton from the ARO Complex Dynamics and Systems Program for his ongoing and continuous support. We would also like to thank Dr Reza Malek-Madani at the ONR for his continued support.

References

[1] C. Cafaro, W. M. Lord, J. Sun, and E. M. Bollt. Causation entropy from symbolic representations of dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(4):043106, 2015.

[2] A. Sudu Ambegedara, J. Sun, K. Janoyan, and E. Bollt. Information-theoretical noninvasive damage detection in bridge structures. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 26(11):116312, 2016.

[3] J. Sun and E. M. Bollt. Causation entropy identifies indirect influences, dominance of neighbors and anticipatory couplings. *Physica D: Nonlinear Phenomena*, 267:49–57, 2014.

[4] J. Sun, C. Cafaro, and E. M. Bollt. Identifying the coupling structure in complex systems through the optimal causation entropy principle. *Entropy*, 16(6):3416–3433, 2014.

[5] J. Sun, D. Taylor, and E. M. Bollt. Causal network inference by optimal causation entropy. *SIAM Journal on Applied Dynamical Systems*, 14(1):73–106, 2015.

THE CONTINUOUS CONFIGURATION MODEL: A NULL FOR COMMUNITY DETECTION ON WEIGHTED NETWORKS

John Palowitch, Shankar Bhamidi, Andrew B. Nobel

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Despite the prevalence of weighted networks in nature, some community detection methods do not easily extend to handle edge weights. One reason for this may be that generative models crucial to community detection methodology, like the configuration model [2] or the degree-corrected stochastic block model [3], are for binary networks. In this work, we introduce an analogue to the configuration model for weighted networks. Our "continuous" configuration model generates a community-less weighted network with given expected degrees *and* expected strengths (node edge-weight sums). Explicitly, let $\mathbf{d}$ and $\mathbf{s}$ be given vectors of degrees and strengths, and A, W the (random) adjacency and weight matrices. We show that under the model, for all nodes i, j,

$$\mathbb{P}(A[i,j] = 1) \propto \mathbf{d}(i)\mathbf{d}(j) \quad \text{and} \quad \mathbb{E}(W[i,j]) \propto \mathbf{s}(i)\mathbf{s}(j).$$

These relationships reflect and extend the 1st-order properties of the configuration model to edge weights. Therefore, we propose the continuous configuration model as a natural null for community detection on heterogeneous weighted networks, much as the standard configuration model has been for unweighted networks [5].

With an explicit, generative null model, diverse avenues for community detection on weighted networks become available. We present a method called Continuous Configuration Model Extraction (CCME), featuring a core iterative procedure with hypothesis tests under the continuous configuration model. CCME is in the style of existing testing-based methods for binary networks based on the standard configuration model [4, 6]. We prove a central limit theorem for edge weight sums under the continuous configuration model, which facilitates a closed-form approximation to p-values inherent to testing algorithm.

Another important facet of our work is the use of a *weighted* stochastic block model (WSBM). Though some stochastic block models with edge weights have been proposed [1], we give a new model that is strength-corrected as well as degree corrected. We employ our WSBM in a theoretical analysis of the continuous configuration model and CCME. We prove that, under standard assumptions, communities from the WSBM are high-probability "fixed points" of CCME: the method recovers them as statistically significant node sets. Importantly, our result allows for network sparsity near the detectability limit, on par with recent consistency analyses for binary networks [7].

We also feature the WSBM in an empirical study of CCME and competing methods. Combining the WSBM and the continuous configuration model, we provide a novel method for simulating weighted networks with both communities and "background" nodes not significantly connected to any community. We also extend the WSBM to involve overlapping communities. In simulation settings involving these diverse flavors of the WSBM, we find that CCME outperforms competitors that are capable of detecting overlapping and background nodes. We also apply CCME and competing methods to real-world weighted networks from various domains, showing that communities found with CCME suggest intuitive insights about the natural systems.

References

[1] C. Aicher, A. Z. Jacobs, and A. Clauset. Learning latent block structure in weighted networks. *Journal of Complex Networks*, page cnu026, 2014.

[2] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *European Journal of Combinatorics*, 1(4):311–316, 1980.

[3] A. Coja-Oghlan and A. Lanka. Finding planted partitions in random graphs with general degree distributions. *SIAM Journal on Discrete Mathematics*, 23(4):1682–1714, 2009.

[4] A. Lancichinetti, F. Radicchi, J. J. Ramasco, S. Fortunato, et al. Finding statistically significant communities in networks. *PloS One*, 6(4):e18961, 2011.

[5] M. E. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[6] J. D. Wilson, S. Wang, P. J. Mucha, S. Bhamidi, A. B. Nobel, et al. A testing based extraction algorithm for identifying significant communities in networks. *The Annals of Applied Statistics*, 8(3):1853–1891, 2014.

[7] Y. Zhao, E. Levina, and J. Zhu. Consistency of community detection in networks under degree-corrected stochastic block models. *The Annals of Statistics*, pages 2266–2292, 2012.

EXTRACTING NEIGHBORHOOD STRUCTURE FROM VERY LARGE DNA GRAPHS

C. Titus Brown, Dominik Moritz, Michael P. O'Brien, Felix Reidl, Blair D. Sullivan

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We describe a method for efficiently computing a hierarchy of r-dominating *graphs* that summarizes the neighborhood structure of a DNA sequence graph at multiple resolutions. This is currently being applied to the problem of metagenome species binning.

Motivation

Metagenomics is the analysis of microbial communities through shotgun DNA sequencing, which randomly samples many subsequences from the genomic DNA of each microbe present in the community. A common problem in metagenomics is the reconstruction of *individual* microbial genomes from the mixture. Typically this is done by first running an "assembly" algorithm that reconstructs longer linear regions based on a graph of the sampled subsequences [2], and then binning the resulting contigs together using compositional analysis of the assembly.

Our novel approach to species extraction extracts bins based on the neighborhood structure of the sequence graph. Because bin extraction involves querying local regions of the graph many times, we build a hierarchical atlas.

Atlas construction

An atlas level consists of a set of nodes, each representing a connected subgraph of G called its *shadow*. The leaves correspond to vertices in an r-dominating set D of G; that is, the minimum distance from every vertex to a member of D is at most r. The shadow of the node for $v \in D$ is the vertices for which v is the closest member of D. Though computing an optimal r-dominating set is NP-hard, we use a linear time approximation algorithm from Dvořák and Reidl [3]. This algorithm uses an efficient path shortcutting subroutine and its approximation factor is low because G has small maximum degree.

The subsequent levels of the atlas are built using a similar process on auxiliary *domination graphs*, which connect members of the previous level's r-dominating set with overlapping shadows. This coarse-graining is designed to respect underlying connectivity.

Querying the atlas

Our atlas is optimized for extracting connected subgraphs of G that contain a set of query vertices and their local neighborhoods. This search proceeds in a top-down manner, building a frontier beginning at the top level and refining progressively towards the bottom. To save space, we store minhash sketches summarizing the shadows with probabilistic coverage guarantees.

Results

To evaluate the construction and search algorithms, we built an atlas for a synthetic data set constructed from known genomes [4], and then used the known genomes to search for genomic bins. Evaluating on a 15-member subset of genomes, we were able to identify atlas nodes containing genome bins with a median sensitivity of 85% and a median specificity of 90%. We are now focusing on engineering the atlas building approach to scale to a 60 million node experimental data set.

Future applications

The atlas may have other genomic applications, for both Genome Wide Association Studies for genetic traits and the analysis of 3-D physical contact maps of chromosomes. Our methodology also applies to neighborhood querying and extraction more generally. The atlas construction is efficient on graphs of *bounded expansion*, which include many real-world networks [1]. We also have implemented the search subroutines and shadow sketching to allow adaptation to specific domain use cases.

References

[1] E. D. Demaine et al. Structural sparsity of complex networks: Random graph models and linear algorithms. *CoRR*, abs/1406.2587, 2014.

[2] J. Pell et al. Scaling metagenome sequence assembly with probabilistic de bruijn graphs. *PNAS*, 109(33):13272–13277, 2012.

[3] F. Reidl. Structural sparseness and complex networks. 2016. Aachen, Techn. Hochsch., Diss., 2015.

[4] M. Shakya et al. Comparative metagenomic and rrna microbial diversity characterization using archaeal and bacterial synthetic communities. *Environ. microbiol.*, 15(6):1882–1899, 2013.

AN17 INVITED PRESENTATION 5: DANIEL SPIELMAN, YALE UNIVERSITY

(Spirit of Pittsburgh A – 3rd floor)

Laplacian Matrices of Graphs: Algorithms and Applications

The Laplacian matrices of graphs arise in fields including machine learning, computer vision, optimization, computational science, and of course network analysis. We will explain what these matrices are and why they arise in so many applications. In particular, we will show how Laplacian system solvers can be used to quickly solve linear programs arising from natural graph problems.

We then will survey recent progress on the design of algorithms that allow us to solve these systems of linear equations in nearly linear time. We will focus on the role of graph sparsification and the recent discovery that it can be used to accelerate Gaussian elimination.

POSTER ABSTRACTS

(in numerical order)

EIGENVALUES FOR RESILIENCE ANALYSIS OF BACKBONE NETWORKS

Egemen K. Çetinkaya and Tristan A. Shatto
Missouri University of Science and Technology

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

A graph spectrum is the eigenvalues and their multiplicities, and graph energy is the sum of eigenvalues. We analyze spectra and energy of backbone networks under different targeted attack scenarios. Our results indicate that while the relative cumulative frequencies of eigenvalues merge to 0, the energy of networks show decreasing and increasing behavior for node and link attacks.

Graph Spectra and Energy

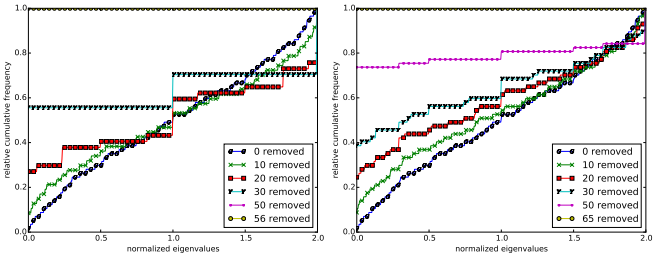
Different data structures can represent connectivity of a network. Let $G = (V, E)$ be an unweighted, undirected graph with n vertices and l edges. $A(G)$ is the symmetric adjacency matrix with no self-loops. The Laplacian matrix of G is: $L(G) = D(G) - A(G)$ where $D(G)$ is the diagonal matrix of node degrees, $d_{ii} = \deg(v_i)$. Given the degree of a node is d_i, the normalized Laplacian matrix $\mathcal{L}(G)$ is:

$$\mathcal{L}(G)(i,j) = \begin{cases} 1, & \text{if } i = j \text{ and } d_i \neq 0 \\ -\dfrac{1}{\sqrt{d_i d_j}}, & \text{if } v_i \text{ and } v_j \text{ are adjacent} \\ 0, & \text{otherwise} \end{cases}$$

Eigenvalues are the roots of the characteristic polynomial. The set of eigenvalues $\{\lambda_1, \lambda_2, \ldots, \lambda_n\}$ together with their multiplicities define the *spectrum*. Graph energy, $\mathcal{E}$, is the sum of absolute values of its eigenvalues [3]. Given an adjacency matrix of a graph, $A(G)$, the graph energy is $\mathcal{E}_A(G) = \sum_{i=1}^{n} |\lambda_i(A)|$ [3]. The graph energy of the Laplacian matrix, $L(G)$, is $\mathcal{E}_L(G) = \sum_{i=1}^{n} |\lambda_i(L) - 2l/n|$ [3]. Lastly, given the normalized Laplacian graph, $\mathcal{L}(G)$, its energy is $\mathcal{E}_{\mathcal{L}}(G) = \sum_{i=1}^{n} |\lambda_i(\mathcal{L}) - 1|$ [1].

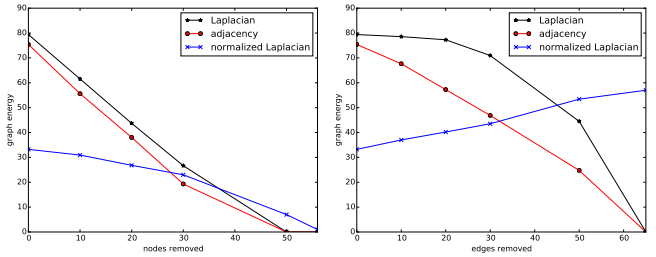
Analysis

We use a realistic dataset of five backbone networks (shown only for the Internet2 fiber-optic network) that are geographically located in the US [2]. The first step of our analysis includes removal of nodes and links based on betweenness centrality. Second, we calculate the spectra and energy as the nodes and links are removed.



(a) Internet2 node removal (b) Internet2 link removal

Figure 1: Spectra of Internet2 backbone network



(a) Internet2 node removal (b) Internet2 link removal

Figure 2: Energy of Internet2 backbone network

Spectra of the Internet2 network are shown in Fig. 1. As nodes and links are removed, eigenvalues merge to a value of 0 [4]. The graph energy of the Internet2 backbone network as nodes and links are removed is shown in Fig. 2. The energy of the graph decreases as nodes are removed. For the link removal scenarios, while the adjacency and Laplacian energy levels decrease, the energy for the normalized Laplacian increases. In conclusion, deterministic eigenvalues are useful in evaluating resilience of graphs.

References

[1] M. Cavers, S. Fallat, and S. Kirkland. On the normalized Laplacian energy and general Randić index R_{-1} of graphs. *Linear Algebra and its Applications*, 433(1):172–190, 2010.

[2] E. K. Çetinkaya et al. Multilevel Resilience Analysis of Transportation and Communication Networks. *Telecommunication Systems*, 60(4):515–537, December 2015.

[3] X. Li, Y. Shi, and I. Gutman. *Graph Energy*. Springer New York, 2012.

[4] T. A. Shatto and E. K. Çetinkaya. Spectral Analysis of Backbone Networks Against Targeted Attacks. In *IEEE DRCN*, 2017.

CONVERGENCE OF THE SPECTRAL RADII FOR RANDOM DIRECTED GRAPHS WITH COMMUNITY STRUCTURE

David Burstein

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

The spectral radius of the adjacency matrix can impact the dynamics in genetic, epidemiological, and biological neural networks. Since observations from real world networks are limited, we want to study the distribution of the spectral radius on a collection of plausible networks using random graph models. And while much literature has focused on the distribution of the spectral radius for undirected random graph models, very little work has explored the analogous problem on directed or asymetrically weighted graphs. Consequently, we address this gap by providing concentration bounds and asymptotics to the spectral radii distribution for a generalization of the directed Chung-Lu random graph model that allows for community structure.

Background

In the directed Chung-Lu random graph model, two nodes share a directed edge with probability proportional to the product of the expected out-degree and in-degree of the two corresponding nodes. And even though the choice of the Chung-Lu random graph may appear arbitrary, this model outputs realizations that emulate many features commonly found in empirically observed networks, such as degree heterogeneity, and is analytically tractable. Even so, deriving spectral results for any directed random graph model poses practical challenges, as the adjacency matrix is no longer symmetric. We circumvent this difficulty by employing a path counting argument as for *any* adjacency matrix A, the average number of cycles of length r and the number of paths of length r provide lower and upper bounds to the rth power of the spectral radius, $\rho(A)^r$, for any choice of $r \in \mathbb{N}$. Deriving concentration bounds on the number of paths and cycles yields the following result.

Theorem 1. *For the Chung-Lu random graph model with lists of the expected in and out degree for each node,*$(\mathbf{a}, \mathbf{b})$*, let* S *be the expected number of edges in the graph. Then it follows that if* $\frac{\mathbf{a}\cdot\mathbf{b}}{S} \to \infty$ *or the maximum of the probabilities that two nodes share an edge approaches* 0*, then with high probability* $\frac{\rho(A)}{\frac{\mathbf{a}\cdot\mathbf{b}}{S}} \to 1$.

Recall that the number of paths of length r provides an upperbound to $\rho(A)^r$. As part of the proof strategy for Theorem 1, we show that for a strategic choice of r, with high probability that the number of paths of length r is bounded above by $C(\frac{\mathbf{a}\cdot\mathbf{b}}{S})^r$, where $C^{\frac{1}{r}} \approx 1$. Note that C could be rather large. Additionally, as the number of paths (or cycles) bounds the spectral radius, we also attain concentration bounds on the spectral radius that apply to networks of finite size as illustrated in Figure 1.

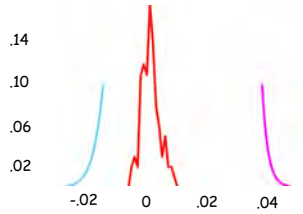


Figure 1: Concentration bounds of the spectral radius for Chung-Lu random graph model with 600 nodes. The red curve is an empirical pdf for the relative error between the observed spectral radius and the quantity $\frac{\mathbf{a}\cdot\mathbf{b}}{S} \approx 161$. The magenta and blue curves are one-sided upper bounds for the likelihood (y-axis) that the spectral radius deviates more than the relative error listed on the x-axis.

Extensions

Path counting also yields spectral radii bounds for a generalization of the Chung-Lu model, where nodes belong to communities and each node has an expected number of incoming and outgoing connections with all nodes in a given community. And even though counting paths in this generalized model becomes much more difficult, as the probability an edge exists now also depends on the community membership of the two nodes, we illustrate how to derive concentration bounds for the number of paths and cycles using the norm of a matrix product. Furthermore, we demonstrate when the spectral radius converges to the spectral radius of an $m^2 \times m^2$ matrix, where there are m communities in the network. Since the spectral radius of the adjacency matrix can influence the dynamics in a network and very little work has considered the spectral radii distribution for directed random graphs, our novel bounds on the spectral radii provide an important tool in the analysis of real world *directed* networks.

THE RELATION BETWEEN ARCHITECTURAL AND FUNCTIONAL CONNECTIVITY IN THE CEREBRAL CORTEX

Paulina Volosov

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

The extent of the relation between architectural and functional connectivity in the cerebral cortex is a question which has attracted much attention in recent years. Neuroscientists frequently use the functional connectivity of neurons, i.e. the measures of causality or correlations between the neuronal activities of certain parts of a network, to infer the architectural connectivity of the network, which indicates the locations of underlying synaptic connections between neurons. Architectural connectivity can be used in the modeling of neuronal processing and in the forming of conjectures about the nature of the neural code. These two types of connectivity are by no means identical, and no one-to-one correspondence or mapping exists from one to the other. In particular, certain trivial measures of functional connectivity, such as correlations, give rise to an undirected network, while synaptic architectural connectivity is always directed. Nevertheless, architectural connectivity can be inferred from functional connectivity, and this work is one attempt to determine how to do so. We will begin by examining different statistical measures of functional connectivity, and, in particular, by determining what directed measures can be employed, for example mutual information, which are better suited for the investigation than mere correlations of firing rates or neuronal voltages. Additional work will involve analyzing the neuronal network structure, looking especially at the incidence matrices representing both types of connectivity with the intention of establishing how one depends on the other. This can be achieved by studying the structure of these matrices through tools including low-rank decomposition and spectral properties.

GENERIC STEADY STATE BIFURCATIONS IN HOMOGENEOUS COUPLED CELL NETWORKS AND RELATED EQUIVARIANT DYNAMICS

Sören Schwenker

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Dynamical systems with an underlying network structure arise frequently in the sciences (e.g. neuroscience) as well as in applications (e.g. electrical circuits). We are especially interested in the class of homogeneous coupled cell networks (cf. Figure 1) for which Rink and Sanders [4] have proved a strong connection to monoid equivariant dynamics. We use this interplay to investigate generic steady state bifurcations – qualitative changes in the steady states due to variation of a parameter – in such networks by means of equivariant dynamics.

Abstract

Rink, Sanders and Nijholt have provided numerous results on dynamics in homogeneous coupled cell networks (cf. [1, 2, 3, 4, 5]). Using their methods, one can present network ODEs as (sub-systems of) systems that are symmetric with respect to a monoid representation. As one was mostly concerned with group representations before this draws attention to very exciting generalizations of representation theory and equivariant dynamics.

We investigate this generalization and especially the classification of generic steady state bifurcations which could up to now only partly be realized: either in the direct context of the networks or in a special case of the monoid representation. We aim at generalizing this last statement to include more general representations. Therefore, we investigate arbitrary representations of arbitrary monoids and examine the generic bifurcation behaviour of equivariant systems. In order to do so, we employ methods from equivariant bifurcation theory of groups on the one hand and the theory of monoid representations on the other. This allows us to extend the bifurcation result to arbitrary monoid representations. Returning to the network context this type of result has the advantage of providing information on the dynamics of classes of networks rather than individual ones as the results hold for all those networks connected to the same monoid representation.

Example

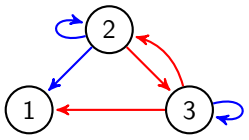


Figure 1: Homogeneous coupled cell network.

The network in Figure 1 induces the ODE system

$$\begin{aligned} \dot{x}_1 &= f(x_1, x_2, x_3) \\ \dot{x}_2 &= f(x_2, x_2, x_3) \\ \dot{x}_3 &= f(x_3, x_3, x_2) \end{aligned} \tag{1}$$

where $x_i \in \mathbb{R}$ and the first component denotes internal dynamics. It is equivariant with respect to the symmetries

$$(x_1, x_2, x_3) \mapsto (x_2, x_2, x_3) \text{ and } (x_1, x_2, x_3) \mapsto (x_3, x_3, x_2)$$

meaning that the right hand side commutes with their application. The symmetries generate a monoid and describe a 3-dimensional representation of it which decomposes as the direct sum of three 1-dimensional subrepresentations. Assuming that system (1) depends on a parameter $\lambda \in \mathbb{R}$ and that it possesses the steady state $(0, 0, 0)$ for all values of λ we search for steady state bifurcations. We find that these generically occur in the direction of the 1-dimensional subrepresentations.

References

[1] E. Nijholt, B. Rink, and J. Sanders. Center manifolds of coupled cell networks. 2016.

[2] E. Nijholt, B. Rink, and J. Sanders. Graph fibrations and symmetries of network dynamics. *Journal of Differential Equations*, 261(9):4861–4896, 2016.

[3] E. Nijholt, B. Rink, and J. Sanders. Projection blocks in homogeneous coupled cell networks. *Dynamical Systems*, 32(1):164–186, 2016.

[4] B. Rink and J. Sanders. Coupled cell networks and their hidden symmetries. *SIAM Journal on Mathematical Analysis*, 46(2):1577–1609, 2014.

[5] B. Rink and J. Sanders. Coupled cell networks: Semigroups, lie algebras and normal forms. *Transactions of the American Mathematical Society*, 367(5):3509–3548, 2015.

POWER NETWORK SCIENCE

David J. Hill, Tao Liu and Yue Song

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

In this work we extend power system theory from a network science viewpoint, which exploits the features of power networks. It appears that the whole of the nearly 100-year old subject of power systems can be rewritten in the network-based paradigm to provide new insights.

Power network modeling

Consider a power system consisting of generator nodes $\mathcal{V}_G$, load nodes $\mathcal{V}_L$ and transmission lines $\mathcal{E}$. The network-reduced model was commonly adopted in earlier results. It eliminates the load nodes via Kron reduction and usually leads to a full graph connecting the generators only. This process reduces the dimension but breaks the original power network structure. Also, the virtual lines created by Kron reduction have considerable resistance representing the effects of loads, which brings difficulties in designing a Lyapunov function for stability and control analysis. To avoid this problem, Bergen and Hill [1] first proposed the *structure-preserving model*

$$M_i\ddot{\theta}_i + D_i\dot{\theta}_i = P_i - \sum\nolimits_{(i,j)\in\mathcal{E}} b_{ij}\sin(\theta_i-\theta_j),\ i\in\mathcal{V}_G$$
$$D_i\dot{\theta}_i = P_i - \sum\nolimits_{(i,j)\in\mathcal{E}} b_{ij}\sin(\theta_i-\theta_j),\ i\in\mathcal{V}_L$$

where $\theta_i, P_i, M_i, D_i, b_{ij}$ denote node angle, power injection, generator inertia, damping constant and line capacity, respectively; and the resistance of physical line can be reasonably neglected. Further, this model can be regarded as a complex dynamical network with the original network structure and heterogeneous node dynamics. So it is a natural model to develop *power network science* [2]. Some recent results have followed this direction [3, 4].

Power network stability

The mainstream approaches for power system stability are node-based, e.g., to study stability by constructing a proper Lyapunov function. The history of these approaches dates back to the early 20th century in Russia and extensive investigations have been conducted [5]. However, the node dynamics evolves via the underlying power network. The role of network topology is of importance but has not drawn enough attention [2]. We shed new light on the instability mechanism of power systems by taking the network-based viewpoint. We reveal that the small-disturbance angle stability can be indicated by the Laplacian matrix of the so-called power flow graph describing the power flows over the network [6]. We also establish matrix conditions in terms of the critical lines to check stability and instability type for all equilibria. Moreover, we apply cutset properties to explain instability phenomenon in large-disturbance scenario [7].

Power network control

With the growing penetration of renewable energy, the control paradigm of power systems is experiencing a profound evolution. The control task is shifting from generator side to demand side, and the physical power network has stronger interaction with the communication network among control devices. A network science view will facilitate the control problems of such complex cyber-physical networks. We have been working towards a distributed non-disruptive demand-side control framework to quickly regulate system dynamics after contingency [8].

Conclusion

We propose the subject of power network science as a product of network science concepts and more theoretical ideas in power system stability and control.

References

[1] A. R. Bergen and D. J. Hill, "A structure preserving model for power system stability analysis," *IEEE Trans. Power App. Syst.*, vol. PAS-100, no. 1, pp. 25–35, Jan 1981.

[2] D. J. Hill and G. Chen, "Power systems as dynamic networks," in *Proceedings of IEEE ISCAS 2006*.

[3] F. Dörfler, M. Chertkov, and F. Bullo, "Synchronization in complex oscillator networks and smart grids," *Proc. Nat. Acad. Sci.*, vol. 110, no. 6, pp. 2005–2010, 2013.

[4] A. E. Motter, S. A. Myers, M. Anghel, and T. Nishikawa, "Spontaneous synchrony in power-grid networks," *Nature Physics*, vol. 9, no. 3, pp. 191–197, 2013.

[5] H.-D. Chiang, *Direct methods for stability analysis of electric power systems: theoretical foundation, BCU methodologies, and applications*. John Wiley & Sons, 2011.

[6] Y. Song, D. J. Hill, and T. Liu, "Network-based analysis of small-disturbance angle stability of power systems," *IEEE Trans. Control Netw. Syst.*, available online.

[7] Y. Song, D. J. Hill, and T. Liu, "Characterization of cutsets in networks with application to transient stability analysis of power systems," *IEEE Trans. Control Netw. Syst.*, available online.

[8] T. Liu, D. J. Hill, and C. Zhang, "Non-disruptive load-side control for frequency regulation in power systems," *IEEE Trans. Smart Grid*, vol. 7, no. 4, pp. 2142–2153, July 2016.

MODULUS METRICS ON NETWORKS(POSTER)

Nethali Fernando

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Metrics and Ultrametrics

Let X be a set. Then $d : X \times X \to \mathbb{R}$ is a *metric* on X if:

- **Non-negativity:** $d(a, b) \geq 0$ for all $a, b \in X$.
- **Non-degeneracy:** $d(a, b) = 0$ if and only if $a = b$.
- **Symmetry:** $d(a, b) = d(b, a)$ for all $a, b \in X$.
- **Triangle inequality:** For every $a, b, c \in X$: $d(a, b) \leq d(a, c) + d(c, b)$.

If instead of the triangle inequality, d satisfies

$$d(a, b) \leq \max\{d(a, c), d(c, b)\}, \qquad \text{for every } a, b, c \in X,$$

then d is an *ultrametric.* Every ultrametric is a metric.

Classical Metrics on Graphs

Three well-known network metrics are shortest path, effective resistance and the (reciprocal of) minimum cut.

p-Modulus

Modulus is a way to quantify the richness of families of objects on graphs, such as families of walks, trees, cycles etc...Here we focus on families of walks. First we recall its definition. For $1 \leq p < \infty$, the *p-modulus* of a family Γ is

$$\mathrm{Mod}_p(\Gamma) := \inf_{\rho \in \mathrm{Adm}(\Gamma)} \mathcal{E}_p(\rho) = \inf_{\rho \in \mathrm{Adm}(\Gamma)} \sum_{e \in E} \rho(e)^p,$$

where $\rho : E \to [0, \infty)$ is *admissible* for Γ ($\rho \in \mathrm{Adm}(\Gamma)$) if

$$\ell_\rho(\gamma) := \sum_{e \in E} \mathcal{N}(\gamma, e)\rho(e) \geq 1 \qquad \forall \gamma \in \Gamma,$$

here $\mathcal{N}(\gamma, e)$ is the number of times γ crosses the edge e.

Connection to Classical Quantities

In the special case of connecting families $\Gamma(a, b)$ modulus recovers some classical quantities [1]. For instance, 2-modulus coincides with effective conductance, when viewing the graph as an electrical network with edge-conductances equal to σ. Also, 1-modulus recovers min cut, and letting p tend to infinity, the p-th root of p-modulus tends to the reciprocal of shortest-path. In general, p-modulus continuously interpolates between these classical measures.

The Main Theorem

For $1 \leq p < \infty$, let $d_p(a, b) := \mathrm{Mod}_p(\Gamma(a, b))^{-1/p}$. If $G = (V, E)$ is a simple connected graph, then d_p is a metric on V. Moreover, d_1 is an ultrametric. See [2].

Antisnowflaking

Whenever d is a metric on X and $0 < \epsilon < 1$, then d^ϵ is also a metric on X. This is called *snowflaking* the metric. What is the largest exponent t such that d^t is still a metric? We introduce the *antisnowflaking exponent* of a metric d: $\mathrm{ASFE}(d) := \sup\{t \geq 1 : \ d^t \text{ is a metric}\}$

Writing $d_{p,G}$ for our modulus metric, to show the dependence on the graph G, we define

$$s(p) := \inf\{\mathrm{ASFE}(d_{p,G}) : G \text{ connected}\}.$$

Conjecture

We conjecture that, for all $p \in (1, \infty)$,

$$s(p) = \frac{p}{p-1}.$$

We arrived at this conjecture analyzing numerical data done on a set of Erdős Rényi graphs and using our own algorithm for computing modulus.
We are currently working on proving this conjecture. The case $p = 1, 2, \infty$ (appropriately definied) are already established. The result follows if one can show that d_p^q is a metric when q is the Hölder conjugate exponent of p.
We have looked at the family of biconnected graphs, complete graphs and hypercubes and so far the results seem in line with the conjecture. We hope to present these numerical analysis and examples worked on different families of graphs in our poster.

References

[1] N. Albin, M. Brunner, R. Perez, P. Poggi-Corradini, and N. Wiens. Modulus on graphs as a generalization of standard graph theoretic quantities. *Conformal Geometry and Dynamics*, 19:298–317, 2015.

[2] N. Albin, N. Fernando, and P. Poggi-Corradini. Modulus metrics on networks. Preprint.

DISCRETE AND CONTINUUM MODELING OF BIOLOGICAL NETWORK FORMATION

Lisa Maria Kreusser, Peter A. Markowich

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Network formation and transportation networks are fundamental processes in living systems. A new dynamic modeling approach to describe the formation of biological transport networks has recently been introduced by Hu and Cai. They propose a continuum model, based on macroscopic laws, as well as a discrete, purely local dynamic adaptation model. We present an overview of recent analytical and numerical results for these models.

Introduction

Transportation networks are ubiquitous in living systems such as leaf venation in plants, angiogenesis of blood vessels and neural networks which transport electric charge. Biologists, engineers, physicists and computer scientists have expressed great interest in understanding natural networks. One of the main research questions is what are the structural and topological properties of optimal networks, in particular the existence of loops.

Description of the model

Traditionally most of the methodological tools use discrete models, based on mathematical graph theory and discrete energy optimization, where the energy consumption within the network is minimized under the constraint of constant total cost. However, biological systems are continuously adapting their structures to meet the changing metabolic demand. Hu and Cai have recently introduced a new dynamic modeling approach [6] accounting for adaptation of networks to fluctuations in the flow in contrast to considering optimization as a global effect. Central to their discrete model is a purely local dynamic adaptation mechanism based on mechanical rules. In particular, this dynamic adaptation model responds only to local information and can naturally incorporate fluctuations in the flow.

To formulate the discrete model we consider a connected graph and associate each vessel of the graph with a nonnegative conductivity. Assuming that the material cost for the vessel i of the network is proportional to the power C_i^γ of its conductivity C_i for a parameter $\gamma > 0$ we consider the energy consumption as the sum of the kinetic energy of the material flow through the vessels and the metabolic cost of maintaining the network. This energy is constrained by the Kirchhoff law which expresses the conservation of mass.

Besides, one can formulate a continuum model based on macroscopic physical laws. This model was introduced in [5], studied in [1, 3, 4, 2] and consists of a very complex system of nonlinear partial differential equations. Because of its highly unusual coupling this model is also of mathematical interest.

Analysis of the model

Using methods from mathematical and numerical analysis we study the discrete and the macroscopic model and investigate the qualitative properties of network structures. Experimental studies of scaling relations of conductivities of parent and daughter edges in real networks suggest that the choice of the parameter γ is crucial for the resulting network formation [3]. This is also underlined by the analytical and numerical results we have obtained indicating a phase transition behavior at $\gamma = 1$ with a uniform sheet, i.e. the network is tiled with loops, for $\gamma > 1$ and a loopless tree for $\gamma < 1$.

References

[1] G. Albi, M. Artina, M. Foransier, and P. A. Markowich. Biological transportation networks: Modeling and simulation. *Analysis and Applications*, 14(01):185–206, 2016.

[2] G. Albi, M. Burger, J. Haskovec, P. Markowich, and M. Schlottbom. Discrete and continuum modelling of biological network formation. Book chapter, submitted.

[3] J. Haskovec, P. Markowich, and B. Perthame. Mathematical analysis of a pde system for biological network formation. *Communications in Partial Differential Equations*, 40(5):918–956, 2015.

[4] J. Haskovec, P. Markowich, B. Perthame, and M. Schlottbom. Notes on a pde system for biological network formation. *Nonlinear Analysis*, 138:127–155, 2016.

[5] D. Hu. Optimization, adaptation, and initialization of biological transport networks. Notes from lecture, 2013.

[6] D. Hu and D. Cai. Adaptation and optimization of biological transport networks. *Physical review letters*, 111:138701, 2013.

NETWORK INDUCED PHASE-LOCKED PATTERNS OF THE KURAMOTO FLOW ON CUBIC GRAPHS

Yury Sokolov, G. Bard Ermentrout

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

We consider the Kuramoto model on sparse graphs, given by the family of 3-regular graphs. We study network properties that ensure the existence of stable equilibria distinct from the synchronized state. The existence of such stable structures may be a consequence of the dependence of the basin of attraction of synchronized states on the structure of the underlying network.

Abstract

One of the main interests in network dynamics is, under which conditions on the underlying graph, does the system exhibit one or another type of behavior. In particular, during the last several decades significant progress has been made on how the network structure affects the convergence and speed of convergence of coupled phase oscillators to the synchronized state. Fundamental rigorous results have been made for extreme cases, i.e., n-cycles and complete graphs, while different graph-based measures were proposed to capture the convergence to the synchronized state for systems defined on general sparse connected graphs.

We consider a simplified Kuramoto model on a family of 3-regular (cubic) graphs with identical frequencies, and unnormalized coupling, which is one for adjacent nodes in the network. We show that there is a nonempty subset of cubic graphs, so that the model defined on those graphs admits stable equilibria distinct from synchronized state – phase-locked patterns. We derive some conditions on graphs under which the network of coupled phase oscillators converges to phase-locked patterns.

The question we address is dual to the study of synchrony. Probably, some of the tools we have developed can be applied in the analysis of the synchronized state on sparse graphs.

NS17 Abstract #31 for

"Clustering Techniques for a Network Derived from Voting"

By Richard Burkhart, Ph.D.

Proportional representation is simple when candidates are identified by group affiliation, but complex when it is voters, not political parties, who rank or rate candidates. Traditional algorithms are ad hoc, offering little insight. Yet clusters of voters may be identified from voting patterns, even when clusters overlap and some voters are left unclustered. We present a clustering algorithm using network science, tested on ballot sets from 38 districts, applying techniques of discrete, continuous, and combinatorial optimization to graphs with weighted edges and vertices. The vertices are mean rating vectors of combined ballots, based on identical top candidates, with vertex correlations specifying edge weights. A cluster is a fuzzy set of vertices with membership from correlation with the cluster mean rating vector. Our non-linear global optimization iterates from diverse initial cluster sets, like the clustering coefficient result, using a damped Newton method, while merging strongly overlapping clusters. Branch and bound is applied to find the best match among sets of candidates to a cluster set, maximizing the sum of cluster-averaged candidate ratings, with proportionality enforced by a soft penalty function, where each cluster is represented fractionally by its best rated elected candidates. Clustering reveals hidden issues of complexity, while yielding more proportional results and bringing the old field of voting algorithms into the realm of modern network science.

REVISITING POWER-LAW DISTRIBUTIONS IN SPECTRA OF REAL WORLD NETWORKS

Nicole Eikmeier, David F. Gleich

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

By studying a few hundred real world graphs up to a few hundred thousand vertices, we find a large body of empirical evidence that most real world graphs have a statistically significant power-law distribution with a cut-off in the singular values of the adjacency matrix and eigenvalues of the Laplacian matrix in addition to the commonly conjectured power-law in the degrees. The observed scaling law in the cut-off enables us to compute only a subset of the spectra of large networks, enabling testing graphs with up to tens of millions of vertices and billions of edges, where we find that those too show evidence of statistically significant power-laws in the spectra. More details about the data and experiments can be found in our paper [1].

Abstract

Power-laws have long been studied in real-world networks such as web-crawls and online social networks. Among other reasons, the purpose of studying these power-law distributions is to generate realistic synthetic network models and to establish theory about why various algorithms work better than expected in networks of this type. While there have been a number of results and findings about power-laws (and the absence thereof!) based on real-world data, synthetic methods, and relationships between eigenvalues and degrees, these studies are often limited to a few small examples. Our specific interest is to investigate power-laws in the singular values of the adjacency matrix, but in the course of this, we also revisit many of these findings with the goal of providing new guidance on the presence of power-laws in three features of real-world networks:

1. the degrees;
2. the singular values of the adjacency matrix;
3. the eigenvalues of the Laplacian matrix;

For our study we considered real-world networks from the Stanford project, Facebook, and various other sources (all data is publicly available) up to a few hundred thousand vertices where we could compute the exact eigenvalues. For comparison we also included a number of network models. In total we studied over $5,000$ distributions. We fit each of the distributions to a power-law, suppling a cutoff in the tail, which gives the size of the distribution included in the power-law, and a test of significance, to gauge the reliability of the results. A power-law is *significant* if it passes this test.

The most interesting result is that power-laws in the singular values appear more consistently than in the degree distribution. Furthermore, a significant power-law distribution in the degrees means there is a high probability for a significant power-law distribution in the singular values of the adjacency matrix and the eigenvalues of the Laplacian matrix. The converse does not hold. The exponents of the power-law distributions are much larger than previously observed, ranging between $2 - 10$. We find a surprising direct relationship between the power-law in the degree distribution and the power-law in the eigenvalues of the Laplacian that was theorized in simple models but is extremely accurate in practice.

We observe a scaling law for the size of the tail equal to $n^{2/3}$ for the degrees and Laplacian eigenvalues and between $n^{2/3}$ and $n^{1/2}$ for the singular values. This allows us to investigate a number of larger networks (up to 65 million vertices) by only considering the top values. These networks are up to 100 times larger than those we used to make our observations in the first part of our study, and include network data-dumps in addition to crawled networks. We find these too have significant power-law distributions in their adjacency singular values, which is consistent with those found on our smaller networks.

This finding is descriptive, but understanding the structure of real world networks allows us to take advantage of inherent structure for faster computation in a variety of settings. In particular, we suspect the results of the reliable power-law in the singular values to be a useful property for characterizing the extremely fast convergence of many matrix-based algorithms on these types of networks.

References

[1] N. Eikmeier and D. F. Gleich. Revisiting power-law distributions in spectra of real world networks. Accepted at KDD2017.

IDENTIFYING TRADE COMMUNITIES IN THE GLOBAL CO2 SUPPLY CHAIN NETWORK USING SINGULAR VALUE DECOMPOSITION

Supun Perera, Somwrita Sarkar, Michael Bell

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

By applying the Singular Value Decomposition (SVD) technique to the global CO2 supply chain dataset presented in Davis et al. [2], we seek to identify the closely knit CO2 trade clusters within the global CO2 supply chain network. In particular, we separately consider the following trade segments (each as a separate supply chain network); (1) Extraction - Production (E-P), (2) Production - Consumption (P-C), and (3) Extraction - Consumption (E-C).

Background

Singular value decomposition

Work by Sarkar et al. [1, 3] has revealed that SVD of a graph matrix, followed by clustering of a reduced space representation, can be used to identify community structures without imposing any a priori assumptions on properties of the communities, such as pre specifying the number or the size. The algorithm uses the largest gaps between the singular values as a heuristic to identify the optimal number of modules in the network. Usually, in a network which has community structure, a small number of singular (or eigen) values will be sharply separated from the bulk distribution of eigenvalues [1].

Global CO2 supply chain network

Conventional production-based accounts of CO2 emissions only represent a single point in the supply chain of fossil fuels, which may have been extracted elsewhere and may be used to provide goods or services to consumers elsewhere [2]. In this regard, it is important to identify the original sources of fossil fuels and the ultimate destinations of goods and services reliant on these fuels, i.e. the entire global supply chain of CO2 emissions. In this study, we use the publicly available trade dataset presented by [2], which spans across the global supply chain of CO2 emissions and tracks global CO2 emissions from the points of Extraction (of all fossil fuels), Production (of goods embodying emissions), and Consumption (of goods embodying CO2 in all industry sectors) for 112 countries.

Results and discussion

E-P: Both extracting and producing countries were found to have five distinct clusters with minimum overlaps. For example, both Australia and New Zealand were found to belong into the same cluster for both extraction and production. On closer analysis, it is evident that both these countries rely heavily on Middle East for extractions. Similarly, both these countries significantly produce to Japan.

E-C: The extracting countries indicated 3 clusters (i.e. countries in each of these clusters extract for almost the same countries). However, high levels of overlap was observed across the clusters, indicating that many countries which extract belong to multiple trade clusters which include different consuming countries. The consuming countries indicated 5 clusters. Much less overlap was observed across these clusters, indicating that many consuming countries belong to clear trade clusters which include unique extracting countries.

P-C: Both producing and consuming countries were found to have two clusters with high levels of overlaps. This indicates that many countries which consume belong to multiple trade clusters which include different countries producing. Similarly, the producing countries belong to clusters which include different countries consuming these products.

Acknowledgements

This work has been funded by the Australian Research Council (ARC) under grant DP140103643.

References

[1] Somwrita Sarkar and Andy Dong. Community detection in graphs using singular value decomposition. *Physical Review E*, 83(4):046114, 2011.

[2] Steven J Davis, Glen P Peters, and Ken Caldeira. The supply chain of co2 emissions. *Proceedings of the National Academy of Sciences*, 108(45):18554–18559, 2011.

[3] Somwrita Sarkar, James A Henderson, and Peter A Robinson. Spectral characterization of hierarchical network modularity and limits of modularity detection. *PloS one*, 8(1):e54383, 2013.

CROSS-VALIDATION ESTIMATE OF THE NUMBER OF CLUSTERS IN COMMUNITY DETECTION

Tatsuro Kawamoto, Yoshiyuki Kabashima

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Community detection is a coarse-graining process of network data, and the assessment of the coarse-grain level is its crucial step. Here we propose [2] principled, scalable, and widely applicable assessment criteria to determine the number of clusters in community detection based on the leave-one-out cross-validation (LOOCV) estimate of the edge prediction error. We also compare the performance of the LOOCV estimates with other popular criteria.

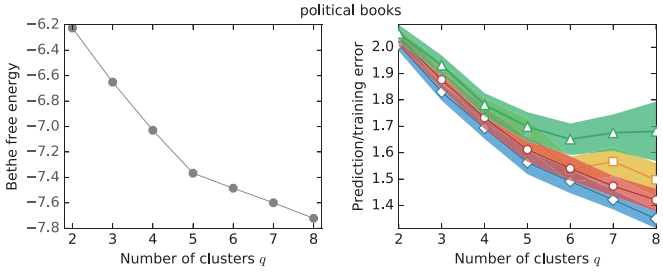


Figure 1: Bethe free energy and cross-validation errors of a real-world network data.

Bayesian inference and model selection criteria

The Bayesian inference using the so-called stochastic block model is one of the popular methods in community detection. While it is computationally infeasible to perform the inference exactly, the EM algorithm with belief propagation [1] is known as the fast and accurate approach even in the case of sparse networks.

When we want to detect communities from a given network data, we need to choose the number of communities. The simplest way to determine the number of communities in the Bayesian framework is to measure the negative marginal log-likelihood, or equivalently, the free energy; when the free energy saturates as we increase the number of communities, the most parsimonious model should be chosen. While this criterion works when the model we assume is consistent with the generative model, in practice, it overfits very often. While the use of the BIC-like criteria can be a choice (e.g., [4, 3]), we introduce to use another well-accepted principle for model selection which is based on the prediction error.

Cross-validation error using belief propagation

One of the advantages of using the prediction errors is that the model we use does not need to be consistent with the generative model of the actual network. While the cross-validation estimate is a standard approach for measuring the prediction errors, it has both conceptual and computational problems when it is naively applied to the inference using the stochastic block model. However, when the LOOCV is considered and the belief propagation is used, we show that both of these problems can be solved. We can conduct the model assessment very efficiently and we confirm that the performance is indeed reasonable for both synthetic and real-world networks (e.g., Fig. 1).

Among the prediction errors that we consider, we cannot generally conclude which one is superior to others theoretically. Instead, we derive a generic inequality among the prediction errors and a formal relation between the prediction errors and the Bethe free energy. Furthermore, when the network is actually generated by the stochastic block model, we show that one of them achieves the information-theoretical limit of detectability.

References

[1] A. Decelle, F. Krzakala, C. Moore, and L. Zdeborová. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E*, 84(6):066106, Dec. 2011.

[2] T. Kawamoto and Y. Kabashima. Cross-validation model assessment for modular networks. *arXiv preprint arXiv:1605.07915, to appear in Sci. Rep.*, 2017.

[3] M. E. J. Newman and G. Reinert. Estimating the number of communities in a network. *Phys. Rev. Lett.*, 117:078301, Aug 2016.

[4] T. P. Peixoto. Parsimonious module inference in large networks. *Phys. Rev. Lett.*, 110:148701, 2013.

CORE DETECTION: SIFTING THROUGH THE JUNK IN GRAPH MATCHING

Vince Lyzinski, Daniel L. Sussman

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

In multiple graph inference often only a fraction of the vertices possess a true match across networks, and many graph matching algorithms do not identify truly matched vertices after aligning the networks. Herein, we present a procedure for detecting correctly matched vertices after the networks have been aligned. We theoretically establish the effectiveness of our procedure in a general bivariate random graph model. These theoretical results are corroborated in both simulated and real data experiments.

Introduction and Background

Given graphs $G_1 = (V_1, E_1)$ and $G_2 = (V_2, E_2)$—with resp. adjacency matrices A and B—the *graph matching problem* (GMP) seeks to minimize $\|AP - PB\|_F$ over permutation matrices $P \in \mathcal{P}$. The GMP has been extensively studied in the literature, with numerous application areas; see, for example, [1]. More recently there has been a flurry of activity tackling the related problem of graph *matchability* [2]: Given a latent alignment ϕ between V_1 and V_2, can graph matching uncover ϕ in the presence of shuffled vertex labels? To model this latent alignment, we define the correlated Erdős-Rényi model.

Definition: *For $R, \Lambda \in [0,1]^{n\times n}$, we say (G_1, G_2)—with resp. adjacency matrices A and B—are R-correlated Erdős-Rényi(Λ) random graphs (abbreviated* $\mathrm{CorrER}(\Lambda, R)$*) if: marginally $G_1, G_2 \sim \mathrm{ER}(\Lambda)$, i.e., $A_{i,j}, B_{i,j} \sim Bern(\Lambda_{i,j})$ independently across i and j; and for each $\{i,j\} \in \binom{V}{2}$, the correlation $corr(A_{i,j}, B_{i,j}) = R_{i,j} \geq 0$.*

Subsequent to the problem of matchability is the problem of match detectability: after matching, can we successfully determine which vertices were correctly aligned by the GM algorithm. Moreover, in applications it is often the case that only a fraction of the vertices in G_1 possess a latent matched pair in G_2, with the remaining vertices having uncorrelated connectivity. We model this by setting $R = R_c \oplus \mathbf{0}_{n_j}$ where $n = n_c + n_j$ and $R_c \in [0,1]^{n_c \times n_c}$ and $\mathbf{0}_{n_j}$ is the $n_j \times n_j$ matrix of all zeros. We call the first n_c vertices the core and the remaining n_j vertices the junk and will refer to this as the core-junk ER model.

Core Detection

After matching core-junk ER graphs, there remains the issue of determining which vertices were, in fact, core vertices. In this model, assuming $R \geq 0$, a natural statistic for testing whether a vertex is a properly aligned core vertex is a GM analogue of Mantel's test statistic, namely $T^*(\cdot) := T(v, A, B, P^*) = \frac{\Delta_v(P^*) - \mathbb{E}_P \Delta_v(P)}{\sqrt{\mathrm{Var}_P \Delta_v(P)}}$, where $\Delta_v(P) = \|(AP - PB)_{v,\bullet}\|_1$, $\mathbb{E}_P$ and Var_P denote the expectation and variance of $\Delta_v(P)$ with respect to uniform sampling of P over all permutation matrices, and $P^* \in \mathrm{argmin}_{P \in \Pi(n_c+n_j)} \|AP - PB\|_F$.

Intuitively, if v is a properly matched core vertex then $\Delta_v(P^*)$ should be significantly smaller—due to the correlation structure—than the number of errors induced by a randomly chosen permutation. If v is a junk vertex or a misaligned core vertex then, even with the small amount of correlation induced by P^*, we expect $\Delta_v(P^*)$ to be closer to $\mathbb{E}(\Delta_v(P^*))$ than in the core setting. In both the core and junk cases, $\sqrt{\mathrm{Var}_P \Delta_v(P)}$ effectively normalizes for the potentially varied degree distributions. Rather than formalizing a hypothesis testing procedure using $T^*(\cdot)$—which necessitates estimating the critical region for $T^*(\cdot)$—we will instead use $|T^*(\cdot)|$ to order the vertices' likelihood of being in the core, with those having higher values of $|T^*(\cdot)|$ more likely to be in the core.

This intuition bears out in both theory and practice: under mild assumptions in the core-junk ER model (satisfied, in one simple example, if $\Lambda = p\vec{1}\vec{1}^T$ and n_c and R are sufficiently large), it holds that $\min_{v \in \mathrm{core}} T(v, A, B, Q) > \max_{u \in \mathrm{junk}} T(u, A, B, Q)$ with high probability for any oracle labeling Q (i.e., $Q = I_{n_c} \oplus Q'$). These theoretic results are corroborated by excellent performance on simulated networks and real data experiments on Twitter data.

References

[1] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. *International Journal of Pattern Recognition and Artificial Intelligence*, 18(03):265–298, 2004.

[2] V. Lyzinski and D. L. Sussman. Graph matching the matchable nodes when some nodes are unmatchable. *arXiv preprint arXiv:1705.02294*, 2017.

EFFICIENT LIKELIHOOD-BASED NETWORK CLUSTERING

Alan Ballard, Marcus Perry

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Clustering networks by maximizing likelihood produces clusters of quality similar or superior to modularity and allows for the statistical testing of clustering significance. However, identifying optimal clusterings based on likelihood quickly becomes computationally prohibitive as the network size grows large. At each proposed re-clustering, a change-in-loglikelihood must be calculated and an accept/reject decision made. While the previous method required the entire network's edge set to be read for each proposed re-clustering, we provide theorems that identify the portions of the network that affect the change-in-loglikelihood given a proposed re-clustering and reduce the new data requirement to just the set of edges connected to the vertex proposed for re-clustering. Further efficiency in computation is achieved by a streamlining of the change-in-loglikelihood formula.

PEELING BIPARTITE NETWORKS FOR DENSE SUBGRAPH DISCOVERY

Ahmet Erdem Sarıyüce, Ali Pınar
Sandia National Laboratories

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

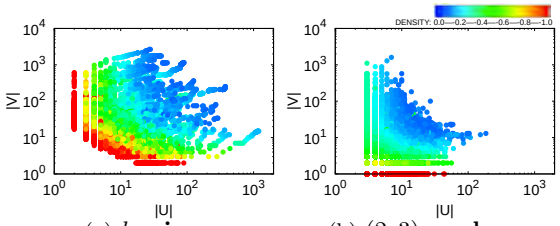
Finding dense bipartite subgraphs and detecting the relations among them is an important problem for affiliation networks that arise in a range of domains, such as social network analysis, word-document clustering, internet advertising, and bioinformatics, to name a few. However, most dense subgraph discovery algorithms are designed for classic, unipartite graphs. Subsequently, studies on affiliation networks are conducted on the co-occurrence graphs (e.g., co-authors and co-purchase networks), which projects the bipartite structure to a unipartite structure by connecting two entities if they share an affiliation. Despite their convenience, co-occurrence networks come at a cost of loss of information and an explosion in graph sizes, which limit the quality and efficiency of solutions. We study the dense subgraph discovery problem on bipartite graphs. We define a framework of bipartite subgraphs based on the butterfly motif (2,2-biclique) to model the dense regions in a hierarchical structure. We introduce efficient peeling algorithms to find the dense subgraphs and build relations among them. Experiments show that we can identify much denser structures compared to the state-of-the-art techniques on co-occurrence graphs. Our algorithms are also memory efficient, since they do not suffer from the explosion in the number of edges of the co-occurrence graph.[1]

Problem and Challenges

Our aim is to find many, if not all, dense regions in bipartite graphs and determine the relations among them by using peeling algorithms. A common practice in the literature for working with bipartite graphs has been creating co-occurrence (projection) graphs. Although the projection enables the use of well-studied unipartite graph mining algorithms, it has significant drawbacks:

- **Information loss and ambiguity:** Bipartite graphs comprise one-to-many relationship information, but this information is reduced to pairwise ties when projected to a weighted or unweighted unipartite form. Those pairwise ties are treated independently, which distorts the original information. In addition, projections are not bijective irrespective of the projection technique being used, which creates ambiguity.
- **Size inflation:** Each affiliation in the bipartite network with degree d_i results in a d_i-clique in the projected graph. Thus, the number of edges in the projected graph can be as many as $\sum_{v \in V} \binom{d_v}{2}$, whereas it is only $\sum_{v \in V} d_v$ in the bipartite network, where V is the set of affiliations. Increase in the number of edges degrades the performance and also artificially boosts the clustering coefficients and local density measures in the projected graph.

Given the drawbacks of projection approaches, we work directly on the bipartite graph to discover the dense structures.



(a) k-**wing** (b) $(2,3)$-**nucleus**

Figure 1: *Dense subgraph profiles for the* IMDb *network. Each dot is a bipartite subgraph, the density, $|E|/(|U| \cdot |V|)$, is color coded and $|U|$ and $|V|$ are given on the x- and y-axis. Wing decomposition algorithm results in 36 bipartite subgraphs with ≥ 0.9 edge density that have at least 10 vertices in each side. Other algorithms working on projections cannot report any bipartite subgraph in that quality.*

Contributions

- **k-tip and k-wing bipartite subgraphs:** We survey attempts to define higher-order structures in bipartite graphs, and use the *butterfly* structure (2,2-biclique) as the simplest super-edge motif. Building on that, we define the *k-tip* and *k-wing* subgraphs based on the involvements of vertices and edges in butterflies, respectively.
- **Extension of peeling algorithms:** We introduce peeling algorithms to efficiently find all the k-tip and k-wing subgraphs. Our algorithms are inspired by the degeneracy based decompositions for unipartite graphs.
- **Evaluation on real-world data:** We evaluate our proposed techniques on real-world networks. Figure **??** gives a glance of results on the IMDb movie-actor with 1.6M vertices and 5.6M edges.

[1] This abstract is based on our paper available on arXiv:1611.02756

INFORMATION DIFFUSION IN COMMON-INTEREST SOCIAL NETWORKS

Rashad Eletreby, Osman Yağan
Carnegie Mellon University
{*reletreb,oyagan*}*@andrew.cmu.edu*

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Common-interest social networks constitute a class of social networks where two individuals are connected if they share a common interest *and* they are socially connected. We model these networks by a composite random graph and derive necessary and sufficient conditions for network-wide information diffusion to be possible. Several future applications of the model will be discussed including inference of missing links in social networks.

Common-interest social networks

A common-interest relationship between two individuals manifests from their selection of common items from a pool of available interests and hobbies [3]. We model these relationships by a *general random intersection graph* (a.k.a. inhomogeneous random key graph [2]), where each of the n nodes is first assigned to one of r possible *classes* with μ_i denoting the probability of a node being class-i. Then, each class-i node selects K_i interests from a large pool (for each $i = 1, \ldots, r$) and a pair of nodes are connected if they have at least one common interest.

This model can be shown to exhibit several characteristics commonly seen in real-world social networks including high *clustering* and *small-world* properties. However, it is clear that people are not necessarily connected with every single individual they have common interests with. In particular, real-world social networks often contain *communities* that form well-connected subgraphs with sparse connection to the rest of the graph. This is often attributed to some form of *homophily*, meaning that individuals tend to be socially connected with a limited number of other individuals who share the same culture, race, etc.

To incorporate these factors into our model, we consider an inhomogeneous Erdős-Rényi (ER) graph [1], where a class-i node and a class-j node are connected with probability α_{ij} independently from everything else. In this setting, we can adjust the $r \times r$ edge probability matrix $\boldsymbol{\alpha}$ suitably to generate a specific structure of the underlying social network formed by multiple communities. For instance, we may set $\alpha_{12} = \alpha_{21} = 0.1$ while $\alpha_{11} = \alpha_{22} = 0.9$ to obtain a network with two relatively well-connected communities formed by class-1 and class-2 nodes, respectively. An added benefit of the inhomogeneous ER model is that, unlike the classical ER graph, it is able to generate scale-free networks with power-law degree distributions [1].

Collecting, our proposed common-interest social network model is formed by the *intersection* of an inhomogeneous random key graph with an inhomogeneous Erdős-Rényi graph. In other words, we propose a model where two individuals are connected if they i) share at least a common interest; *and* ii) have an edge in the inhomogeneous ER model indicating that they have a social tie (e.g., they belong to the same community). In this talk, we will present some recent results concerning the *connectivity* of this intersection model. In particular, we will present a sharp threshold result identifying the *critical* scaling of the parameters involved for the model to be connected almost surely. We discuss various implications of our results including those concerning the feasibility of global information diffusion in the common-interest network. We will also discuss various other applications of the model for future work, particularly on the inference of missing links in social networks.

References

[1] B. Bollobás, S. Janson, and O. Riordan. The phase transition in inhomogeneous random graphs. *Random Structures and Algorithms*, 33(1):3–122, 2007.

[2] O. Yağan. Zero-one laws for connectivity in inhomogeneous random key graphs. *IEEE Transactions on Information Theory*, 62(8):4559–4574, Aug 2016.

[3] J. Zhao, O. Yağan, and V. Gligor. k-connectivity in secure wireless sensor networks with physical link constraints-the on/off channel model. *arXiv preprint arXiv:1206.1531*, 2012.

THE CONTROL OF ARBITRARY SIZE NETWORKS OF LINEAR SYSTEMS VIA GRAPHON LIMITS

Shuang Gao, Peter E. Caines

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

To achieve control objectives for extremely complex and very large scale networks using standard methods is a challenging, if not intractable, task. In this work we propose a novel way to achieve approximate control for such networks by using the theory of graphons and infinite dimensional system theory.

Network Systems and Limit Graphon Systems

Consider an interlinked network S^N of linear dynamical subsystems $\{S_i^N; 1 \leq i \leq N\}$, each with an n dimensional state space. Each subsystem is uniquely associated to a vertex of the N node graph G_N whose undirected edges correspond to the dynamical interactions between the subsystems. We specify the (symmetric) linear dynamics for the network S^N via the equation

$$\dot{x}_t = A_N \circ x_t + B_N \circ u_t, x_t, u_t \in R^{nN}, A_N, B_N \in R^{nN \times nN}, \quad (1)$$

where $A_N = A_N^T$ denotes a (matrix weighted) adjacency matrix of G_N, $B_N = B_N^T$ denotes a linear input-to-state mapping, and $\circ$ denotes the so called averaging operator given by $A_N \circ x = \frac{1}{(nN)} A_N x$. The adjacency matrices can be represented by step functions (see [2]) in the graphon space $\mathbf{G_1^{sp}}$, i.e. the space of symmetric measurable functions $W_1 : [0,1]^2 \to [-1,1]$. Then trajectories of the

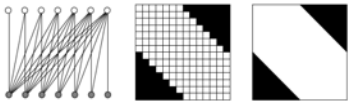


Fig.1 A half-graph, its step function and its graphon limit. (See [2].)

system (1) correspond one-to-one with the trajectories of the system

$$\dot{x}_t^s = A_s^{[N]} x_t^s + B_s^{[N]} u_t^s, \quad x_t^s, u_t^s \in L^2_{step}[0,1], A_s^{[N]}, B_s^{[N]} \in \mathbf{G_1^{sp}},$$

$$[A_s^{[N]} x_s](\alpha) := \int_0^1 A_s^{[N]}(\alpha, \beta) x_s(\beta) d\beta, \quad x_s \in L^2[0,1].$$

$\mathbf{G_1^{sp}}$ is compact under the cut metric [2] and complete under the $L^2[0,1]^2$ metric. Let the graphon sequences $\{A_s^{[N]}\}$ and $\{B_s^{[N]}\}$ be Cauchy sequences of step functions in $L^2[0,1]^2$ with graphon limits A and B (which will then necessarily also be the limits in the cut metric, see [2]). The limit system $(A; B)$ is given by

$$LS^\infty : \dot{x}_t = Ax_t + Bu_t, \quad x_t, u_t \in L^2[0,1], A, B \in \mathbf{G_1^{sp}},$$

where, A and B are graphons, and hence as operators on $L^2[0,1]$ are bounded and hence continous; furthermore, A generates a C_0-semigroup. Specializing the theory in [1] to the case of $L^2[0,1]$ Hilbert state spaces, one can show that the graphon system LS^∞ has a unique mild solution $x \in C([0,T]; L^2[0,1])$ for any $x_0 \in L^2[0,1]$ and any $u \in L^2[0,T; L^2[0,1]]$ (see [1]) .

Define $W_T = \int_0^T e^{At} B B^T e^{A^T t} dt$ as the *controllability Gramian operator*, then the criterion for exact controllability (see [1]) of the system LS^∞ is that, for all $h \in L^2[0,1]$, $\langle W_T h, h \rangle \geq c_T \|h\|^2$, where $c_T > 0$.

The Graphon Control Strategy

(1) Consider the general control problem of steering the states of each member of a sequence S of network systems $\{S^N; 1 \leq N \leq \infty\}$ to each of a sequence x_T of desired states $\{x_T^N; 1 \leq N \leq \infty\}$, where it is assumed that S converges to some limit system LS^∞ and x_T to some x_T^∞. (2) Specify the corresponding control problem CP^∞ for LS^∞ on $L^2[0,1]$ and choose a tolerance $\varepsilon > 0$. (3) Find the control law u^∞ for CP^∞. (4) Then Theorem 1 below and the convergence of the x_T sequence yield N_ε such that $x_T^N(u^N)$ is within ε of x_T^∞ and of x_T^N for all $N \geq N_\varepsilon$.

Theorem 1 *Assume* $(A; B)$ *and* $(A_s^N; B_s^{[N]})$ *are exactly controllable, then there exist controls* u^∞ *and* u^N *such that*

$$\|x_T^\infty(u^\infty) - x_T^N(u^N)\|_2 \leq \|A_\Delta^N\|_2 \|B\|_2 \int_0^T e^{T-\tau}(T-\tau) \cdot \|u_\tau^\infty\|_2 d\tau + \|B_\Delta^N\|_2 \int_0^T e^{(T-\tau)\|A_s^{[N]}\|_2} \cdot \|u_\tau^\infty\|_2 d\tau$$

where $x_T^\infty(u^\infty) = x_T^\infty$, $A_\Delta^N = A - A_s^{[N]}$ *and* $B_\Delta^N = B - B_s^{[N]}$.

References

[1] A. Bensoussan, G. Da Prato, M. C. Delfour, and S. Mitter. *Representation and control of infinite dimensional systems*. Springer Science & Business Media, 2007.

[2] L. Lovász. *Large networks and graph limits*, volume 60. American Mathematical Soc., 2012.

COMMUNITY IDENTIFICATION USING SPIKING NEURAL NETWORKS

Kathleen E. Hamilton, Neena Imam, Travis S. Humble

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Community detection and the related problem of graph partitioning is a robust area of research in many scientific disciplines. This has lead to the proliferation of a wide range of methods and algorithms [1, 6, 7]. Our approach to community detection for an undirected unweighted graph, is inspired by the use of Hopfield recurrent neural networks to the task of graph partitioning [5] and is a clustering based method which incorporates the generation and interpretation of spiking data.

A graph $\mathcal{G}(V, E)$ is mapped to a system of artificial neurons $\mathcal{S}(N, W)$. Each vertex of the graph $(v_i \in V)$ is mapped to a parameterized spiking neuron $(n_i(t_R, v_{th}, \tau) \in N)$, and each undirected graph edge in (E) is mapped to a positively weighted, symmetric pair of synaptic connections $(w_{ij}(s_W) \in W)$. Through the careful tuning of neuron parameters and the use of external stimuli, we extract information about community structure from the similarity in neuron firing patterns. Our approach has been applied to small $(n < 50)$ graphs with two or more clearly delineated communities, such as the barbell graph [8] and the relaxed caveman graph [1].

We use systems of homogeneous neurons, described by the leaky-integrate and fire model of neuron dynamics [2], and simulated using the Python library Brian2 [3]. The nonlinear equation of motion for the j-th neuron is, $\dot{v_j}(t) = (1/\tau)(v_j(t) - I_j(t))$ and the neuron fires a spike when $v(t) > v_{th}$. The spiking pattern generated by neuron n_j is dependent on an electrical term: $I_j(t) = \sum_{t_f} s_W \delta(t - t_f) + I_j^{ext}(t)$, which incorporates arrival of synaptic impulses and an external driving current.

To identify the communities in a barbell graph, a pair of external driving currents were applied to neurons (n_i, n_j) in different communities: $I_i^{ext}(t) = A\sin(\omega t)$, $I_j^{ext}(t) = -A\sin(\omega t))$. The positive amplitude driving was large enough to cause neuron n_i to spike, and depending on the synaptic weight s_W would subsequently cause its neighboring neurons to spike. The negative amplitude driving is necessary to inhibit the spiking of neuron n_j and to inhibit the spread of the spike pattern.

This mapping and driving approach shows it is possible to distinguish between two communities in small graphs using spiking data. It requires careful tuning of all neuron parameters and is less effective when applied to larger graphs with more than 2 communities or fuzzy communities. For these graphs, driving with square pulses is efficient stimuli [4] when the mapping is modified in include positive and negative weighted synapses.

Acknowledgements

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

References

[1] S. Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[2] W. Gerstner and W. M. Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge University press, 2002.

[3] D. F. Goodman and R. Brette. The Brian simulator. *Frontiers in neuroscience*, 3:26, 2009.

[4] K. E. Hamilton, N. Imam, and T. S. Humble. Community detection with spiking neural networks for neuromorphic hardware. Neuromorphic Computing Symposium: Architectures, Models, and Applications, Knoxville, TN, 2017. in preparation.

[5] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the theory of neural computation*, volume 1 of *Santa Fe Institute studies in the sciences of complexity*. Addison-Wesley, 1991.

[6] F. D. Malliaros and M. Vazirgiannis. Clustering and community detection in directed networks: A survey. *Physics Reports*, 533(4):95–142, 2013.

[7] M. T. Schaub, J.-C. Delvenne, M. Rosvall, and R. Lambiotte. The many facets of community detection in complex networks. *Applied Network Science*, 2(1):4, 2017.

[8] D. J. Watts. Networks, dynamics, and the small-world phenomenon 1. *American Journal of sociology*, 105(2):493–527, 1999.

SOFT PAIRWISE CONSTRAINTS IN CLUSTERING AND COMMUNITY DETECTION

Nathan D. Cahill, Alexander Cloninger

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

In real-world situations where graph-based models are used for clustering or community detection, expert users or analysts frequently have knowledge in the form of pairwise *must-link* or *cannot-link* constraints that could be injected into the clustering process. We develop a framework for incorporating this information into the multiway normalized cut cost [2] in a soft manner that handles conflicts, and we extend this framework to Newman-Girvan modularity [1] by illuminating the connection between modularity and normalized cuts. Finally, we illustrate the impact of soft constraints in both image segmentation and in the analysis of brain network connectivity.

Background

Consider an undirected weighted graph $\mathcal{G} = (V, \mathcal{E})$ with vertex set $V = \{v_1, \ldots, v_n\}$ and edge set $\mathcal{E} \subseteq V \times V$ that we wish to partition into k disjoint subgraphs $\mathcal{G}_i = (V_i, \mathcal{E}_i)$, $i = 1, 2, \ldots, k$, where $\bigcup_{i=1}^{k} V_i = V$. Let $\mathbf{X} \in \mathbb{R}^{n \times k}$ be a partition indicator matrix so that $X_{i,j} = 1$ if $v_i \in V_j$ and $\mathbf{X}_{i,j} = 0$ otherwise. A partitioning cost used widely in computer vision is the multiway normalized cut [2]:

$$\mathrm{NCut}_{\mathbf{W}}(V_1, \ldots, V_k) = \tfrac{1}{k}\mathrm{tr}\Big(\mathbf{X}^{\mathbf{T}}\mathbf{L}\mathbf{X}\big(\mathbf{X}^{\mathbf{T}}\mathbf{D}\mathbf{X}\big)^{-1}\Big) \quad , \quad (1)$$

where $\mathbf{W}$, $\mathbf{D}$, and $\mathbf{L}$ are the weighted adjacency, degree, and Laplacian matrices of $\mathcal{G}$. We show that (1) can be written equivalently by:

$$\mathrm{NCut}_{\mathbf{W}}(V_1, \ldots, V_k) = \tfrac{1}{k}\mathrm{tr}\left(\hat{\mathbf{X}}_\ell^{\mathbf{T}}\mathbf{L}\hat{\mathbf{X}}_\ell'\Big(\hat{\mathbf{X}}_\ell^{\mathbf{T}}(\mathbf{D}-\mathbf{K})\hat{\mathbf{X}}_\ell'\Big)^{-1}\right) \quad (2)$$

for any $\ell = 1, \ldots, k$, where $K_{i,j} = d_i d_j / \mathrm{Vol}(V)$, d_i is the degree of V_i, $\mathrm{Vol}(V) = \sum_{v_i \in V} d_i$, and where $\hat{\mathbf{X}}_\ell \in \mathbb{R}^{n \times (k-1)}$ is the matrix formed by removing the ℓ^{th} column of $\mathbf{X}$.

Incorporating Pairwise Constraints

If $\mathcal{S} \subseteq V \times V$, $\mathbf{\Theta}$ is a matrix of nonnegative weights, and $\mathbf{e}_j$ is the j^{th} column of the identity matrix, we define the following *weighted potential matrix*, to encode different types of pairwise constraints:

$$\mathbf{Q}_{\mathcal{S},\mathbf{\Theta}} = \sum_{(v_i,v_j)\in\mathcal{S}} \theta_{i,j}\big(\mathbf{e}_i\mathbf{e}_i^{\mathbf{T}} + \mathbf{e}_j\mathbf{e}_j^{\mathbf{T}} - \mathbf{e}_i\mathbf{e}_j^{\mathbf{T}} - \mathbf{e}_j\mathbf{e}_i^{\mathbf{T}}\big) \quad . \quad (3)$$

Now, suppose that $\mathcal{M}$ and $\mathcal{C}$ are sets of ordered pairs of vertices for which must-link (ML) and cannot-link (CL) constraints are desired, respectively, and that the desired strengths of these pairwise constraints are given by the matrices $\mathbf{\Gamma}$ and $\mathbf{\Xi}$. The NCut objective (2) can be modified to simultaneously incorporate both sets of constraints by adding $\mathbf{Q}_{\mathcal{M},\mathbf{\Gamma}}$ to $\mathbf{L}$ and $\mathbf{Q}_{\mathcal{C},\mathbf{\Xi}}$ to $\mathbf{D} - \mathbf{K}$.

Connection to Modularity

The formulation (2) allows us to see an immediate connection to the Newman-Girvan modularity [1]. According to the Newman-Girvan null model, a random graph is constructed so that the probability of an edge between vertices v_i and v_j is $K_{i,j}/\mathrm{Vol}\,(V)$. Hence, (2) is normalizing the partitioning cost on $\mathcal{G}$ by the expected partitioning cost on the null model.

The beauty of this interpretation of normalized cuts is that *soft cannot-link constraints on the original graph are re-interpreted as soft must-link constraints on the null model.* The same idea can be used to define a soft pairwise-constrained version of modularity maximization.

Computing and Illustrating Solutions

It is straightforward to form spectral relaxations for soft pairwise-constrained versions of normalized cuts and modularity, allowing optimal values of each of these functions to be approximated through the solution of generalized eigenvalue problems. We will show how the resulting approximation algorithms can be applied to semi-supervised image segmentation and to the investigation of functional brain networks.

References

[1] M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(026113):1–15, 2004.

[2] S. X. Yu and J. Shi. Multiclass spectral clustering. In *Proc. International Conference on Computer Vision*, pages 313–319. IEEE, 2003.

OPTIMAL DEPLOYMENT OF RESOURCES FOR MAXIMIZING IMPACT IN SPREADING PROCESSES

Andrey Y. Lokhov, David Saad

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

The effective use of limited resources for controlling spreading processes on networks is of prime significance in diverse contexts, ranging from the identification of "influential spreaders" for maximizing information dissemination and targeted interventions in regulatory networks, to the development of mitigation policies for infectious diseases and financial contagion in economic systems. Most existing algorithms for optimal resource allocation in spreading processes are based on topological characteristics of the underlying network and aim to maximize impact at infinite time. Solutions for these optimization tasks that are based purely on topological arguments are not fully satisfactory; in realistic settings the problem is often characterized by heterogeneous interactions and requires interventions in a dynamic fashion over a finite time window via a restricted set of controllable nodes. The optimal distribution of available resources hence results from an interplay between network topology and spreading dynamics.

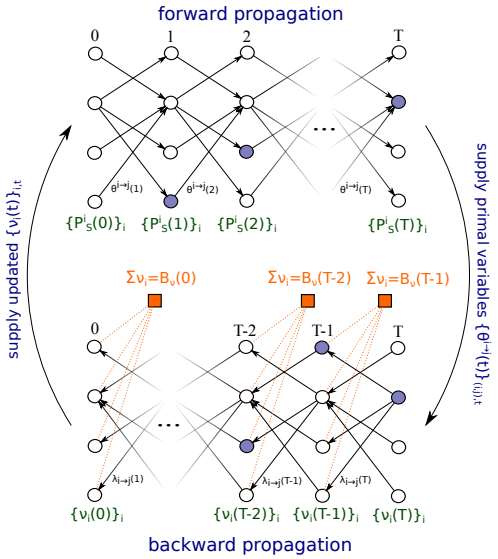


Figure 1: Forward-backward propagation for implementing the constrained dynamic message-passing algorithm.

In this contribution [3], we introduce a new probabilistic targeting formulation which incorporates the dynamics and encompasses previously considered optimization problems. We show how the resulting set of problems can be addressed as particular instances of a universal analytical framework based on two ingredients: scalable dynamic message-passing equations [1, 2] which allow for an efficient solution of the dynamics, and forward-backward propagation, a gradient-free optimization method inspired by the techniques used in artificial neural networks [4] and implemented on top of the constrained message-passing scheme (see Fig. 1). We demonstrate the efficacy of the method on very large synthetic graphs, as well as on a variety of real-world examples, see Fig. 2 for an illustration.

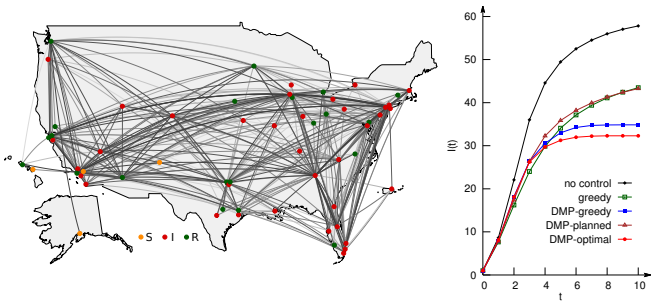


Figure 2: Efficacy of the method applied to the online mitigation of air-traffic mediated epidemic on the real-world network of flights between major U.S. hubs.

References

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

[2] A. Y. Lokhov, M. Mézard, and L. Zdeborová. Dynamic message-passing equations for models with unidirectional dynamics. *Phys. Rev. E*, 91(1):012811, 2015.

[3] A. Y. Lokhov and D. Saad. Optimal deployment of resources for maximizing impact in spreading processes. *arXiv preprint arXiv:1608.08278*, 2016.

[4] D. Saad and M. Rattray. Globally optimal parameters for online learning in multilayer neural networks. *Phys. Rev. Lett.*, 79(13):2578, 1997.

Symmetric and asymmetric coarsening schemes for large-scale networks

Ilya Safro
School of Computing
Clemson University
http://www.cs.clemson.edu/~isafro

A general approach for solving many computational and modeling problems on large-scale networks is through multilevel (also known as multiscale, multiresolution, etc.) algorithms. This approach generally involves coarsening the problem, by producing a sequence of progressively coarser levels (smaller, hence simpler, related problems), anf then recursively using a solution of each coarse problem to provide an initial approximation to the solution at the next-finer level. At each level, this initial approximation is improved by what is generally called "local processing" (LP). This is an inexpensive sequence of short steps, each involving only a few unknowns, together covering all unknowns of that level several times over. Typical examples of LP are few sweeps of relaxation (such as Gauss-Seidel) in the case of solving a system of equations, a few Monte Carlo passes in simulations, or node refinement in partitioning. Following the LP, the resulting approximation may be further improved by one or several cycles, each using again a coarser-level approximation followed by LP, applying them at each time to the residual problem (the problem of calculating the error in the current approximation).

At each level of coarsening one needs to define the set of coarse unknowns. Each coarse unknown is defined in terms of the next- finer-level unknowns (defined, not calculated: they are all unknowns until the coarse level is approximately solved and the fine level is interpolated from that solution). In network problems, each node of the coarse network can represent an aggregate of several fine-level nodes or a weighted aggregate of such nodes, that is, allowing each fine-level node to be split between several aggregates. In the process of defining the set of coarse variables and in constructing an explicit interpolation, it is important to know how "close" two given fine-level variables are to each other at the stage of switching to the coarse level. We need to know, in other words, to what extent the value after the LP of one variable implies the value of the other. If they are sufficiently close, they can, for example, be aggregated to form a coarse variable. Addressing an issue of how to measure the "closeness" between two nodes and how to design an appropriate coarsening is central to many methods and applications.

We will review a class of relaxation-based methods termed *algebraic distance* along with several (a)symmetric coarsening strategies. These measures of closeness define the distance between one node i and a small subset S of several nodes by measuring how well their values are correlated at the coarsening stage, namely, following the LP relaxation sweeps.

An essential aspect of the algebraic distance defined here is that it is a crude local distance. It measures meaningful closeness only between neighboring nodes; the closer they are the less fuzzy is their measured distance. For nodes that should not be considered as neighbors, their algebraic distance just detects the fact that they are far apart; its exact value carries no further meaning. The important point is that this crude local definition of distance is fast to calculate and is all that is required for the coarsening purposes. A similar notion of distance is then calculated at each coarser level.

The coarsening strategies that are based on the proposed measures of closeness can be represented in various forms of algebraic multigrid inspired algorithms. We will present both symmetric and asymmetric coarsening schemes with applications to solvers of Laplacian systems of equations, node immunization, and network ordering. Our novel asymmetric coarsening Laplacian solvers demonstrate a significant improvement in the convergence ratio, and error. The node immunization symmetric coarsening strategy provides efficient and effective heuristics that outperform such solvers as COUENNE and BARON and combinations of several local search methods. The network ordering problems are all different versions of the minimum p-sum problems with applications to compression, cache-friendly layout and clustering.

THE EVOLUTION OF FLOW-BASED HIERARCHY IN NETWORKS

Zizhen Chen, David W. Matula and Eli V. Olinick

SIAM Workshop on Network Science 2017
July 13-14 · Pittsburgh

Extended Abstract

Divisive (top-down) graph-decomposition methods based on edge centralities (e.g., Girvan and Newman 2002 [4], and Fortunato et al. [3]) have been developed to avoid deficiencies associated with agglomerative clustering methods [5]. These community detection methods iteratively identify and remove high centrality edges to produce a hierarchical decomposition of the graph into clusters (connected components).

The divisive algorithm investigated in this paper is based on "maximin" concurrent flow and its dual sparsest cut. Formally, the *maximum concurrent flow problem* (MCFP) is a maximum network flow problem in which every pair of nodes can send and receive flow concurrently. The term *throughput* is defined to be the ratio of the flow supplied between a pair of nodes to the given demand for that pair. The objective of the MCFP is to maximize the throughput, which must be the same for all pairs of nodes, subject to fixed capacity constraints on the edges [1].

A canonical MCFP solution is characterized by a maximal set of "slack" edges with residual flow capacity identifying a partition into connected components of slack edges. The complementary "critical edges" are saturated with flow by any MCFP solution, typically comprising an edge "cut set" forming a bipartition of the graph. The hierarchical MCFP (HMCFP) then further maximizes the common throughput between all node pairs connected by a path of slack edges determining a second throughput level and second set of critical edges bifurcating all the nodes. Iterating further, a series of throughput levels is determined until all edges are critical, yielding a hierarchical stratification portrayed as a *dendrogram* [6].

Real world networks such as the 15-node Florentine Families network [2] are represented as graphs. Taking edge capacities and node pair demands both as unity provides a density based hierarchy by the HMCFP. For the HMCFP at every throughput level, the slack edges at that level identify a partition into component "cluster nodes". The cluster node partition at each level identifies a graph contraction.

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 backbones for the Florentine Families graph HMCF partitions into 3 and 10 parts are shown in Figure 1. The backbones visually display relationships between clusters at each level and introduce a "distance" between clusters.

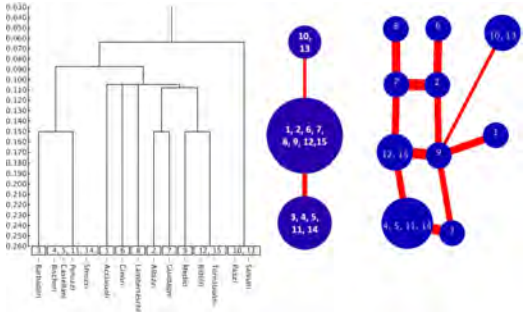


Figure 1: Florentine Dendrogram and two Backbones

References

[1] P.-O. Bauguion, W. Ben-Ameur, and E. Gourdin. Efficient algorithms for the maximum concurrent flow problem. *Networks*, 65(1):56–67, 2015.

[2] C. L. DuBois. UCI network data repository. http://networkdata.ics.uci.edu, 2008.

[3] S. Fortunato, V. Latora, and M. Marchiori. Method to find community structures based on information centrality. *Phys. Rev. E*, 70:056104, Nov 2004.

[4] M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7821–7826, 2002.

[5] C. F. Mann, D. W. Matula, and E. V. Olinick. The use of sparsest cuts to reveal the hierarchical community structure of social networks. *Social Networks*, 30(3):223–234, 2008.

[6] P. Sneath and R. Sokal. *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* Freeman, 1973.

PROCLIVITY PATTERNS IN ATTRIBUTED GRAPHS

Reihaneh Rabbany, Dhivya Eswaran, Christos Faloutsos, and Artur W. Dubrawski

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Many real world applications include information on both attributes of individual entities as well as relations between them, while there exists an interplay between these attributes and relations. For example, in a typical social network, the similarity of individuals' characteristics motivates them to form relations, a.k.a. social selection; whereas the characteristics of individuals may be affected by the characteristics of their relations, a.k.a. social influence. We can measure proclivity in networks by quantifying the correlation of nodal attributes and the structure [1]. Here, we are interested in a more fundamental study, to extend the basic statistics defined for graphs and draw parallels for the attributed graphs.

More formally, an attributed graph is denoted by (A, X); where $A_{n\times n}$ is the adjacency matrix and encodes the relationships between the n nodes, and $X_{n\times k}$ is the attributes matrix –each row shows the feature vector of the corresponding node. Degree of a node encodes the number of its neighbors, computed as $k_i = \sum_j A_{ij}$. We can extend this notion to networks with binary attributes to the number of neighbors which share a particular attribute x, i.e. $k_i(x) = \sum_j A_{ij}\delta(X_j, x)$; where $\delta(X_j, x) = 1$ iff node j has attribute x. Similar to the simple graphs, where the degree distribution is studied and shown to be heavy tail, here we can look at: 1) the degree distributions per attribute, 2) the joint degree distribution of any pair of attributes. Moreover, if we assume $A(x_1, x_2)$ is the induced subgraph (or masked matrix of edges) with endpoints of values (x_1, x_2), i.e., $(A(x_1, x_2))_{ij} = A_{ij}\delta(X_i, x_1)\delta(X_j, x_2)$, then we can study and compare these distributions for the induced subgraph per each pair of attribute values. For example, Figure 1 shows the same trend in the degree distribution of the original graph and three induced subgraphs for CoRA Citations Network [2], with 11,881 papers with 31,482 citations between them, in which a single attribute AI indicates whether topic of the corresponding paper is Artificial Intelligence.

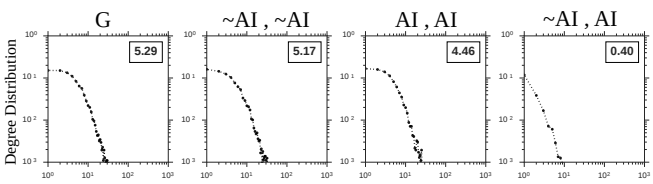


Figure 1: Degree distribution in attribute induced subgraphs.

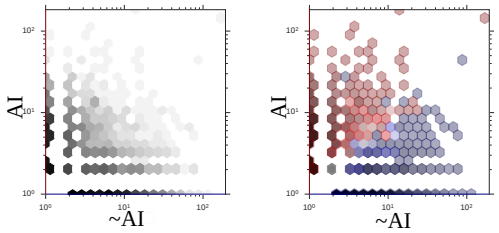


Figure 2: Homophily in degrees: the joint frequency of non-AI neighbors (x-axis), and AI neighbors (y-axis). We can further color divide this distribution based on the value of the node itself, i.e., AI papers are marked by red, and non-AI with blue.

Algebraically, we can compute the degree distributions as the marginals of the adjacency matrix A, i.e., $A\mathbf{1}$. This can be generalized to attributed graphs by considering the matrix multiplication of adjacency matrix A with feature/attribute matrix X, which results in a $n \times k$ matrix AX, in which columns show the degree distribution of nodes for the corresponding attribute value, i.e., number of neighbors of that particular attribute each node has, e.g., number of female friends. In case of two attributes, we can plot the resulting two columns to compare the number of female v.s. male friends per each node. Figure 2 shows such comparison for CoRA, which has a significantly strong proclivity[1] of 0.72, based on the mixing matrix of $\{\{37472, 2380\}, \{2380, 20732\}\}$, due to homophily.

We call AX the "degree matrix", since $(AX)_{ij}$ denotes the number of neighbors that node i has, which have j^{th} attribute, i.e., $(AX)_{ij} = \sum_k A_{ik}X_{kj}$. Here, each column, $(AX)_{:j}$, shows the degree distribution for attribute j, i.e., the number of neighbors which have the j^{th} attribute, per each node; and each row shows the attribute distribution for neighbors of node i, i.e., number of neighbors node i has per each attribute value.

In the same fashion, we study different patterns in attributed networks to reach a better understanding of these ubiquitous datasets which are emerging in diverse domains.

References

[1] R. Rabbany, D. Eswaran, A. W. Dubrawski, and C. Faloutsos. Beyond assortativity: Proclivity index for attributed networks (prone). In *PAKDD*, 2017.

[2] P. Robles-Granda, S. Moreno, and J. Neville. Sampling of attributed networks from hierarchical generative models. In *KDD*, 2016.

MEASURING NOVELTY AND IMPACT WITH EVOLVING HYPERGRAPHS

Feng Shi, James Evans

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

As the complexity of science grows, we are increasingly challenged by the diversity and scale of knowledge required to solve complex research problems. To efficiently navigate through this complex space of knowledge, we need a "map" of the current state of knowledge as well as a computable framework to organize knowledge in a useful way. Here we model scientific and technological knowledge as a complex system, built up from heterogeneous interactions between diverse, differentiated components, and develop a stochastic block model for hypergraphs to describe the evolution of this system quantitatively.

Introduction

Discoveries and inventions are commonly modeled as combinatorial processes of existing knowledge – building on the shoulders of giants. Despite the continuous effort in addressing this myth of successful explorations, little has been achieved. Uzzi *et al.* (Science 2013) attempted to address this myth by showing that papers with both conventional and novel pairwise combinations of journal references are associated with high citation counts. This suggests the question: can we use these traces of knowledge production to effectively guide scientific investigations? Our answer is yes but we need to tackle directly on the content of knowledge – the chemicals, diseases, methods, physical entities, concepts and their relationships studied in those papers. We also need new methods to account for the high-dimensional relationships between those components. We employ a hypergraph framework to examine how scientists and engineers successfully construct novel ideas and objects.

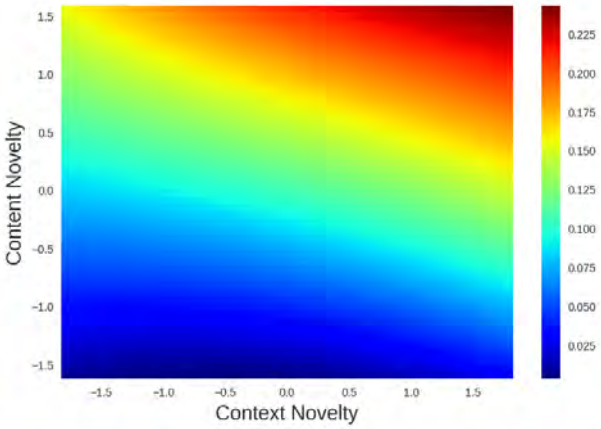
Data and Method

We begin by mapping the complex space of knowledge onto a network. First of all, elementary components of scientific and technological knowledge are identified using community curated ontologies — MeSH terms for biomedical knowledge, and subclasses for patents. We then represent the articles and patents as combinations of those components and also contexts or subfields they draw upon. Mathematically this representation is a hypergraph and combinations of components are hyperedges.

To quantitatively characterize those combinations of contents or contexts, we develop a mixed-membership stochastic block model for hypergraphs, along with a hidden Markov process over the time sequence of hypergraphs to capture their evolution. This model is generative, allowing us to predict future combinations based on those before in a principled, statistical manner.

On the system level, we find large-scale structures and communities in the network. On a finer scale, the hypergraph model can be used to predict promising but unseen combinations (with an AUC$\approx$ 0.9).

The improbability of new combinations predicts success in terms of citations (shown as a heatmap below).



PREDICTING HIGH CENTRALITY VERTICES IN TIME VARYING NETWORKS

S. Sarkar, S.Sikdar, A. Mukherjee, S. Bhowmick

SIAM Workshop on Network Science 2016
July 15-16 · Boston

Introduction

In this paper [1] we present a two-step algorithm for predicting high betweeness and closeness centrality vertices in time-varying networks. In contrast to earlier work [2] on predicting the average betweeness centrality of time varying networks, we predict the exact ids which is more useful in a practical context. For example, if it is known a-priori which vertices would have high centrality in the future time steps, then proper infrastructure can be set up beforehand to utilize these vertices.

Overview of Algorithm

The key idea behind our prediction is to classify networks based on a property, that we term as Core Connectedness.

Consider a graph G with cores numbered from outer to inner(top) respectively as $1, \ldots, K_i, K_{i+1}, \ldots, K_{max}$. We classify the edges as inter-core (intra-core) when the end points are in two different (same) cores. *We define a network to be Core Connected if* $I_{max,i} < \delta_i - J_{max,i} \forall i$. Here $I_{max,i}$ ($J_{max,i}$) is the average inter (intra) core distance between K_{max} and K_i, and δ_i is the average of intra-core within core K_i. We can analytically show that if this condition is maintained then most of the shortest paths pass through K_{max}.

We use the following metrics to identify whether a network is Core Connected; *(i)* Fraction of inter-edges connected to the top core (higher is better), *(ii)* Average density of the non-top cores (lower is better), *(iii)* Density of the top-core (higher is better). We also check a fourth property *(iv)* the overlap of the vertices in the top core at consecutive time steps (higher is better).

In the first step of the algorithm, we use an autoregressive-integrated-moving-average (ARIMA) [1] to predict the Jaccard overlap between the nodes in the top core at a future time step. In the second step once the network is available, we can precisely identify the top central nodes by identifying high degree vertices in the top core. For networks where the four criteria are maintained, our average F-score for prediction is **0.60** (best **0.81**) for closeness and **0.58** (best **0.72**) for betweenness centrality.

[1] A longer version of this work has been submitted to a conference.

Results

Table 1 shows the results of our predictions on a set of real world networks. These are Autonomous Systems(AS (V:7K, E:27K) and CA (V:31K,E:1M), Citation Networks(HT (V:34K, E:4M) and HP(V:27K, E:3M)) and Online social networks(SO (V:2M, E:36M), FW (V:46K, E:2M), SU (V:194K, E:924K).

Each network is classified as a four tuple (column 1) with G representing good and B representing bad. Mean, standard deviation are reported for both prediction error and F-Score. The categories are colored as per the groups they belong. The higher, the number of Gs in the category, the more accurate the prediction results.

Category	Network	CC Prediction (top 10)	F-score e
GGGG	AS	5.69,6.37	0.81,0.06
GGGG	CA	8.76,6.02	0.77,0.08
GGBG	HT	26.96,17.44	0.42,0.35
GGBG	HP	11.64,5.76	0.42,0.33
BBBG	SO	27.96,21.69	0.35,0.26
BBBB	FW	109.90, 92.39	0.24,0.25
BBBB	SU	147.06,106.53	0.02,0.09
Category	Network	BC Prediction (top 10)	F-score
GGGG	AS	6.97,7.68	0.72,0.08
GGGG	CA	9.17,6.47	0.64,0.07
GGBG	HT	20.74,14.86	0.52,0.30
GGBG	HP	14.22,11.23	0.46,0.29
BBBG	SO	26.15,24.72	0.39,0.30
BBBB	FW	56.19, 34.95	0.20,0.19
BBBB	SU	32.58,40.14	0.18,0.21

Table 1: Classification and the prediction performance for real-world networks.

References

[1] G. E. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung. *Time series analysis: forecasting and control*. John Wiley & Sons, 2015.

[2] H. Kim, J. Tang, R. Anderson, and C. Mascolo. Centrality prediction in dynamic human contact networks. *Computer Networks*, 56(3):983–996, 2012.

MINIMIZING CONGESTION IN SUPERMARKETS WITH QUEUING NETWORKS

Fabian Ying, Mason A. Porter, Sam Howison, Mariano Beguerisse-Díaz

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Abstract

Reducing congestion inside stores (e.g., supermarkets) is of great interest to many retailers; congestion affects customer experience, and may delay the fulfillment of online orders. We model stores as planar graphs in which nodes represent zones, and edges connect the nodes of neighboring zones. Customers traverse through the graph via the edges, and they queue to be served at each node. Once they have been served, they visit the next node. This approach allows us to apply standard results from queuing theory to find queue sizes and other quantities of interest.

We analyse how the connectivity of the network affects the total mean queue size Q, our measure of congestion. We also find the network structures that minimize Q under reasonable constraints on the network structure, which gives insight into the store layouts with less congestion. We would like this contribution to be considered for a poster only.

Title: Brain Network Modeling for Epilepsy Based on EEG Signals

Jianzhong Su
Department of Mathematics
University of Texas at Arlington
USA

Honghui Zhang
School of Natural and Applied Science
Northwestern Polytechnical University
China

Ariel Bowman
Department of Mathematics
University of Texas at Arlington
USA

E-mail: Su@uta.edu

Abstract:

Multi-channel Electroencephalography (EEG) signals measure the brain field potential fluctuations on the skull and we can mathematically calculate the electric current density inside the brain by solving an inverse problem. Based on these data, we can build functional network and connectivity of various brain areas during particular tasks of brain. In this talk, we briefly introduce mathematical methods for the EEG source reconstruction problems and discuss some of the applications in finding abnormality in brain activities during seizures of an infant patient with Glucose Transporter Deficiency Syndrome. We will also discuss a dynamic seizure model on a system of ordinary differential equations built on the network topology suggested from the EEG data, and compare the modeled dynamics with experimental data.

ANOMALY DETECTION IN DYNAMIC NETWORKS USING MULTI-VIEW TIME-SERIES HYPERSPHERE LEARNING

Xian Teng and Yu-Ru Lin

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Detecting anomalous patterns from dynamic and multi-attributed network systems has been a challenging problem due to the complication of temporal dynamics and the variations reflected in multiple data sources. We propose a *Multi-view Time-Series Hypersphere Learning* (MTHL) approach that leverages multi-view learning and support vector description to tackle this problem. Given a dynamic network with time-varying edge and node properties, MTHL projects multi-view time-series data into a shared latent subspace, and then learns a compact hypersphere surrounding normal samples with soft constraints. Our approach has several advantages: (1) it preserves the original temporal regularities, (2) it extracts robust representations from multi-view data sources, (3) it produces an optimized hypersphere that allows for effectively distinguishing normal and abnormal cases.

Extended Abstract

Anomaly detection in dynamic network systems has attracted lots of attention in recent years. Most prior works do not take temporal variations into account [1, 3] – they divide streaming data into fixed-length segments and use integrated features as inputs to train models. The integration of attributes might lead to potential loss of temporal information that is critical for anomaly detection. Besides, those techniques primarily focus on single-view data – that is, data captured from a single or homogeneous data source. In addition, many studies are not able to provide a specific representation of normal patterns, which is significant for understanding a system. To deal with these tasks, we propose a novel approach called *Multi-view Time-series Hypersphere Learning* (MTHL). Our framework exploits mutual supportive multi-view time-series, preserves the temporal structure of streaming data, and learns a network's normal patterns by optimizing a hypersphere.

As shown in Figure 1, we consider a dynamic network with a set of attributed vertices and dynamic relationships among them. The network can describe a diversity of

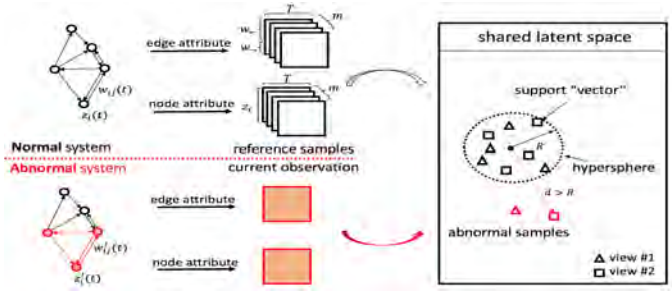


Figure 1: Illustration of our framework.

real-world systems, for instance a city. In this case, the vertices can represent regions with the attributes reflecting their socio-economic factors, and the edges represent transportation flows between regions. An anomalous region can be detected if its time-dependent attributes or relationships deviated from what would normally be expected. **First**, to preserve the temporal variation of multiple attributes, we use multivariate time series representation – a chronologically ordered sequence of feature vectors to capture variation in attribute values. **Second**, inspired by [2], we assume that all multi-view time series share a common latent subspace, and the data from the normal system would be projected into a compact area in the latent space. MTHL learns a hypersphere around the reference set and distinguishes normal and abnormal. Our experiments on synthetic and real-world data demonstrate MTHL's promising results in detecting anomalous events. Moreover, our approach exhibits consistent and good performance in dealing with noises, anomaly pollution in training phase and data imbalance.

References

[1] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *SIGKDD*, pages 1166–1175. ACM, 2014.

[2] S. Li, Y. Li, and Y. Fu. Multi-view time series classification: A discriminative bilinear projection approach. In *CIKM*, pages 989–998. ACM, 2016.

[3] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *SIGKDD*, pages 1176–1185. ACM, 2014.

NETWORK COMPLEXITY, REDUNDANCY AND EFFICIENCY IN THE DEVELOPING HUMAN BRAIN

Catherine Stamoulis

SIAM Workshop on Network Science 2016
July 15-16 · Boston

Summary

Across multiple scales of spatio-temporal variation, the adult human brain is characterized by networks of neuronal ensembles that differentially synchronize their activity in response to cognitive demands and/or external inputs. Topologically, these are networks have small-world and scale free architectures that facilitate optimally efficient processing of neural information. Optimality may be associated with low network redundancy, high efficiency in hierarchical processing and degeneracy. Although a number of studies have shown that small-world networks represent a fundamental aspect of the functional organization of the healthy adult brain, little is known about the functional topologies of brain networks and their properties in the developing brain. In particular, it is unclear how fundamental topological properties of adult brain networks emerge in early life, as the brain undergoes profound structural and functional changes, including selective connection pruning and strengthening, to facilitate increasingly efficient information processing and complex behaviors. To address this question, longitudinally acquired brain data from large cohorts is necessary, particularly given significant variability of neural activity in the developing brain, in part associated with unique experiences in early life.

To robustly characterize dynamically varying neuronal network architerctures during early development, a cohort of 395 healthy infants was measured longitudinally with scalp EEG from 6 to 36 months of life. The dynamic reconfigurations and progressive re-organization of both task-independent and functional networks were estimated using probabilistic directional and non-directional connectivity measures. It is shown that there is a significant re-organization of brain network topologies even in the first year of life, resulting in decreased network redundancy and increased efficiency. However, the hierarchical organization of neural information processing does not emerge until later in life by 36 months of age. Yet, the infant brain may be able to perform the required cognitive tasks. It is shown that this may be in part due to an inherent degeneracy in brain networks that may be in place early in life.

EVIDENCE ACCUMULATION ON NETWORKS

Simon Stolarczyk, Daniel Poll, Zachary Kilpatrick, Krešimir Josić

SIAM Workshop on Network Science 2017
July 13-14 · Pittsburgh

Summary

We study collective decision making on a network using a two-alternative choice task. Typically, group evidence accumulation is modeled heuristically with a coupled drift-diffusion equation. Here we take a more principled Bayesian approach and investigate the behavior of rational agents who can only observe each others actions. Interestingly, this analysis shows that even when no action is observed, information can still be communicated between the agents when their decision thresholds are asymmetric. In recurrent networks the situation become more complex, as agents have to account for correlations in the information received.

Introduction

The two-alternative choice task has been thoroughly studied for a single observer [1]. Here an observer attempts to ascertain the true state of the world, $H \in \{\pm 1\}$, by making sequential noisy observations. An ideal observer sums the resulting log likelihood ratios, until the sum of the evidence reaches one of two pre-determined decision thresholds. Each threshold is identified with a decision.

Model

We extend this single-agent model of evidence integration to a directed network with N agents. Each of the agents is trying to infer the state of the true state of the world, $H \in \{\pm 1\}$. Again, each agent makes independent observations about this state, and accumulates evidence to reach a decision. However, the agents can observe each other's decisions, but have no access to each other's observations. Thus, explicit communication between agents connected by a directed edge occurs only when the agent being observed makes a decision. We describe how an ideal agent integrates information received from other agents combined with its own direct observations to reach a decision.

Results

If agents are biased, they set different evidence thresholds for their decision. We show that in such asymmetric situations if all agents know the evidence thresholds of their neighbors, then agents can obtain information even from observing that their neighbors have *not* made a decision. Even in the simple case when the evidence distributions are discrete and $N = 2$, the evidence gained from observing a neighbor who has not yet made a decision is heavily dependent on the thresholds (See Figure 1).

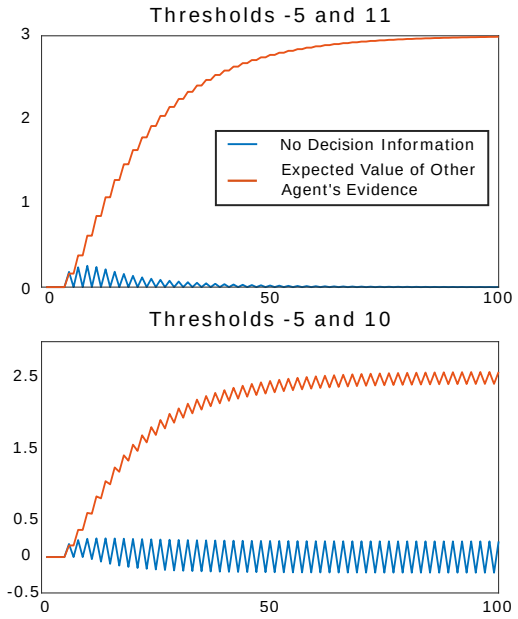


Figure 1: Each plot shows (in blue) how much evidence is gained from knowing a single neighbor has not made a decision for two different sets of thresholds. It also shows (in orange) the expected value of that neighbors evidence given that they have not made a decision.

We derive stochastic differential equations in the continuum limit of many observations. An analysis of these equations, confirmed by simulations shows, that the structure of the network affects the decision proces. Finally, we compare the results to other models such as a coupled diffusion equation [2].

References

[1] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological review*, 113(4):700, 2006.

[2] V. Srivastava and N. E. Leonard. Collective decision-making in ideal networks: The speed-accuracy tradeoff. *IEEE Transactions on Control of Network Systems*, 1(1):121–132, 2014.

HIGHER ORDER STRUCTURE DISTORTS LOCAL INFORMATION IN NETWORKS

Xin-Zeng Wu, Allon G. Percus, Kristina Lerman

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

The information that is locally available to individual nodes in a network may significantly differ from the global information. We call this effect *local information bias*. This bias can significantly affect collective phenomena in networks, including the outcomes of contagious processes and opinion dynamics.To quantify local information bias, we investigate the strong friendship paradox in networks [1], which occurs when a majority of a node's neighbors have more neighbors than it does itself.

Consider a network with a degree distribution $p(k)$ where nodes have a binary attribute x (e.g., active vs inactive). A node is in the paradox regime if attribute values x' for more than half of its neighbors are 1. The global probability of the strong friendship paradox is $P_{>\frac{1}{2}} = \sum_k p(k) f(k)$, the weighted sum of observations by nodes with degree k. Assuming that the degrees of neighbors are independent and identically distributed (iid), the probability of the paradox for each degree class is [2]

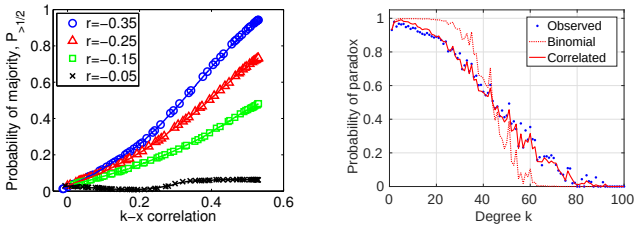
$$f(k) = \sum_{n>\frac{k}{2}}^{k} \binom{k}{n} \mu_x(k)^n [1-\mu_x(k)]^{k-n} \qquad (1)$$

Here $\mu(k)$ is the probability a degree k node has $x = 1$,

$$\mu_x(k) = \sum_{k'} P(x'=1|k') \frac{e(k,k')}{q(k)}, \qquad (2)$$

where $e(k,k')$ is the joint distribution of degrees of linked nodes, and $q(k) = \sum_{k'} e(k,k')$. Fig. 1a shows $P_{>\frac{1}{2}}$ in a power law network for different assortativity r (which is given by $e(k,k')$). Thus, even when few nodes are active (have $x = 1$), many others will observe the "majority illusion" [2], i.e., see that the majority of their neighbors are active.

For the degree version of the strong friendship paradox, we can define an indicator function $x_i = \mathbf{1}_{k_i'>k}$. The node is in the paradox regime if $\overline{x} \equiv \frac{1}{k}\sum_{i=1}^{k} x_i > \frac{1}{2}$. Then Eq. (2) becomes $\mu(k) = \sum_{k'>k} \frac{e(k,k')}{q(k)}$. However, plugging this into Eq. (1), the function $f(k)$ (dotted line in Fig. 1b) does not fit the data. This suggests that the neighbors are *correlated* [3]. Thus, Eq. (1) must be modified to represent



(a) (b)

Figure 1: (a) The global fraction of nodes that have a majority active neighbors in a scale-free network as a function of degree-attribute correlations and assortativity. Only 5% of the nodes are actually active. (b) Comparing observed paradox in a citation network to predictions of the binomial and the correlated model.

a multivariate rather than a single binomial distribution:

$$f(k) = 1 - \Phi\left\{\frac{\frac{1}{2} - \mu_x(k)}{\sigma_x(k)}\right\}, \qquad (3)$$

the variance is $\sigma_x^2(k) = \frac{1}{k}\mu_x(k)[1-\mu_x(k)][1+(k-1)\rho_x(k)]$. The function $\rho_x(k)$ is the degree correlation of two random neighbors x_i and x_j. This has to be derived from the joint degree distribution of a connected ordered triplet of nodes $t(k_i, k, k_j)$. From the information given by $\mu_x(k)$ and $\rho_x(k)$ we can calculate $f(k)$ in high precision (solid line in Fig. 1b).

Our analysis identified certain properties that determine the strength of the paradox in a network: attribute-degree correlation, network assortativity and neighbor-neighbor degree correlation. We also discovered that the neighbor-neighbor degree correlation is significant in real world networks. Understanding how the paradox biases local observations can inform better measurements of network structure and our understanding of collective phenomena.

References

[1] F. Kooti, N. O. Hodas, and K. Lerman. Network Weirdness: Exploring the Origins of Network Paradoxes. In *ICWSM*, 2014.

[2] K. Lerman, X. Yan, and X.-Z. Wu. The "majority illusion" in social networks. *PLoS ONE*, 11(2):e0147617, 2016.

[3] X.-Z. Wu, A. G. Percus, and K. Lerman. Neighbor-neighbor correlations explain measurement bias in networks. 2016.

INFERENCE OF INTERACTION NETWORKS USING CAUSATION ENTROPY

Warren M. Lord, Jie Sun, and Erik M. Bollt
Clarkson Center for Complex Systems Science

SIAM Workshop on Network Science 2016
July 15-16 · Boston

Abstract

Many complex systems of scientific interest are too complicated to derive realistic models of their dynamics from scientic principles, or even determine from these principles which variables interact with each other. An alternative approach, based on empirical processing of experimental observations, is to sample the variables in time and try to infer the dependencies in the form of a directed graph, an approach that we combined with the concept of information to develop Causation Entropy (CSE). We demonstrate the successful application of CSE when exact values of entropy can be calculated (revealing the behavior of toy symbolized stochastic processes under misplaced partitions). We also show the successful use of CSE in conjunction with efficient nonparametric estimation of entropy (simulations of highly nonlinear discrete time dynamical and stochastic systems and an application to collective animal motion using empirically gathered data). We are currently working with neuroscientists at Wake Forest to uncover brain function from fMRI imagery in terms of known anatomical regions. We will discuss how the directed graph of information flows depend on the discretization of time in continuous time systems and also details regarding the definition and estimation of entropy when the samples lie near an underlying measure zero attractor.

NETWORK GIBBS HOMOLOGY AND BETTI NUMBER IDENTIFY NOVEL THERAPEUTIC TARGETS IN EWING SARCOMA

Drew FK Williamson, Jacob G Scott

SIAM Workshop on Network Science 2017
July 13–14 · Pittsburgh, PA, USA

Summary

Many traditional bioinformatics expression analyses discard much of the data inherent in the interactions known to occur between genes and proteins. We harness these data as well as toplogical considerations of protein interaction networks to identify novel therapeutic targets in Ewing sarcoma, a rare tumor characterized by transcriptional dysregulation.

Abstract

Ewing sarcoma (ES) is a deadly and mysterious childhood cancer. Though advances have been made in its treatment, survival has plateaued over the last several decades as insights into the genotype and phenotype of ES cells have yielded disappointingly few viable treatment options [1]. Approximately 85% of ES cases are characterized by a translocation between chromosomes 11 and 22, resulting in a fusion protein known as EWS-FLI. Unfortunately, this fusion proteins structure is not well-conserved and currently undruggable [4]. If we are ever to cure this disease, we must explore genes differentially regulated by EWS-FLI and core pathways that we can perturb to change the ES phenotype.

Aside from the EWS-FLI translocation, ES is a genomically quiet disease. Therefore, our study focuses on changes at the RNA level. RNA-Seq studies by our collaborators uncovered nearly 6,000 genes differentially regulated by EWS-FLI; sitting atop a hierarchy with complex interconnections allows EWS-FLI to create an environment where it is difficult to tease out signal from noise.

We apply a network-based approach that combines protein interaction networks (PIN) and expression-level analysis by superimposing RNA-Seq data onto a PIN and considering this structure as an abstraction of the coupled thermodynamic processes that underlie all actions of a cell. This paradigm allows us to calculate Gibbs free energy which takes into account both the topology of the network and how up- or down-regulated a protein/gene is compared to interactors. We have found robust correlations of PIN Gibbs free energy with disease progression and survival in multiple cancer types [2]. To search for key pathways, we utilize the concept of persistent homologies, features of a surface that can be detected after smoothing out roughness and noise. For the Gibbs free energy landscape, this amounts to identifying the subnetworks that are the primary contributors to a phenotype.

Once the network has been pruned to only the most important constituents, we analyze the network using Betti number to find genes whose removal would most destabilize the network by breaking as many cycles as possible. The more cycles a network has, the greater its robustness: if one link in a cycle is broken or a node removed, the network can reroute around it.

Work published by our collaborators has demonstrated this technique can uncover novel therapeutic targets [3]. This idea is currently being used in phase II trials in Europe to identify personalized drug targets for cancer patients.

We report strong qualitative and quantitative differences in the Gibbs landscapes of relatively treatment resistant and sensitive ES cell lines. The cell lines display major alterations in which biologically relevant subnetworks are upregulated. We identify several novel and several current therapeutic targets using Betti numbers.

References

[1] N. Esiashvili, M. Goodman, and R. B. Marcus. Changes in incidence and survival of ewing sarcoma patients over the past 3 decades: Surveillance epidemiology and end results data. *Journal of pediatric hematology/oncology*, 30(6):425–430, 2008.

[2] E. A. Rietman, J. Platig, J. A. Tuszynski, and G. L. Klement. Thermodynamic measures of cancer: Gibbs free energy and entropy of protein–protein interactions. *Journal of biological physics*, 42(3):339–350, 2016.

[3] E. A. Rietman, J. G. Scott, J. A. Tuszynski, and G. L. Klement. Personalized anticancer therapy selection using molecular landscape topology and thermodynamics. *Oncotarget*, 8(12):18735, 2017.

[4] A. Üren and J. A. Toretsky. Ewing's sarcoma oncoprotein ews–fli1: the perfect target without a therapeutic agent. *Future Oncology*, 1(4):521–528, 2005.