Construction of and efficient sampling from the simplicial configuration model

Construction of and efficient sampling from the simplicial configuration model

Jean-Gabriel Young,[1,*] Giovanni Petri,[2] Francesco Vaccarino,[2,3] and Alice Patania[2,3,†]

[1]*Département de Physique, de Génie Physique, et d'Optique, Université Laval, G1V 0A6 Québec (Québec), Canada*

[2]*ISI Foundation, 10126 Torino, Italy*

[3]*Dipartimento di Scienze Matematiche, Politecnico di Torino, 10129 Torino, Italy*

(Received 29 May 2017; published 22 September 2017)

Simplicial complexes are now a popular alternative to networks when it comes to describing the structure of complex systems, primarily because they encode multinode interactions explicitly. With this new description comes the need for principled null models that allow for easy comparison with empirical data. We propose a natural candidate, the *simplicial configuration model*. The core of our contribution is an efficient and uniform Markov chain Monte Carlo sampler for this model. We demonstrate its usefulness in a short case study by investigating the topology of three real systems and their randomized counterparts (using their Betti numbers). For two out of three systems, the model allows us to reject the hypothesis that there is no organization beyond the local scale.

DOI: 10.1103/PhysRevE.96.032312

Network science's approach to complexity rests onto the tacit hypothesis that the structure of complex systems is reducible to the pairwise interaction of their constituents. It is often a valid premise and, as a result, network science has been extremely successful in, e.g., both predicting [1] and controlling [2] the behavior of complex systems, inferring their function from their structure [3,4], and so on. Networks, however, might not be as ubiquitous as previously thought. It has been shown recently that the structure of a number of complex systems, such as the brain [5,6], protein interactions [7], and social systems [8,9], cannot always be reduced to the sum of pairwise interactions. For these systems, it is now known that network representations can give an incomplete picture: When many-body interactions are broken down into multiple pairwise interactions (cliques), high-order information simply disappears [10].

Simplicial complexes generalize graphs by encoding many-body interactions explicitly; they have hence been proposed as a complementary description of the structure of complex systems [11–14]. Different from hypergraphs, they are equipped with an implicit notion of containment. If nodes $(v_1, \ldots, v_{q+1})$ are involved in a q-dimensional interaction, then it is implicit that all possible lower-dimension interactions involving the same nodes also exist [for example, $(v_1, \ldots, v_q)$ and (v_1, v_3)]. While it might appear constraining, this property actually arises in all systems where interactions are maximal, e.g., in scientific collaborations (largest cohesive group of collaborators) or gene activation pathways (largest group of collectively activated genes). Furthermore, it is found in many processed relational datasets, e.g., in *clique complexes*, obtained by mapping the cliques of networks to simplices [15,16], or in filtered simplicial complexes [13]. Simplicial complexes thus offer a natural and compact description of the structure of complex systems, both when high-order structures are explicitly available or when they are extracted from low-order information.

This application of simplicial complexes has led to promising discoveries: We now better understand, for instance, how to detect large viral recombination events [17], how brain networks reorganize under drugs [18], and how the atomic structure of amorphous solids is hierarchically organized [19]. It has become crucial to establish the statistical significance of these findings, a task for which random null models will be needed. There is already a rich and growing literature on random simplicial complexes and topology, ranging from simplicial generalization of Erdös-Rényi models, amendable to analytical treatment [20,21], to equilibrium formulations of simplicial complex ensembles [10,22], and growth models that reproduce various emergent patterns observed in real systems [23,24]. However, null models—in the sense of network science—are still wanting [25,26].

We address this issue by refining a recently proposed generalization [22] of the (simple) configuration model of network science [25,27,28], which we dub the simplicial configuration model (SCM). Different from Ref. [22], we think of our model as a null hypothesis for real systems; we therefore develop a numerical and statistical toolbox instead of focusing on closed ensemble averages. This entails a number of interesting results: One, we define the first simplicial configuration model able to describe arbitrary complexes, in line with our goal of obtaining a generic null model (Sec. I). Two, we propose and analyze an efficient and rigorous sampling algorithm for this model (Sec. II). Three, we use the model to investigate real datasets and show—now using sound statistical arguments—that the local structure of these systems does not always explain their mesoscale structure (Sec. III). We conclude by listing a few important open problems.

I. SIMPLICIAL CONFIGURATION MODEL

Informally, a labeled simplicial complex K is the high-order generalization of a network. Formally, it is a collection of simplices incident on a node set $V = \{v_1, \ldots, v_n\}$ [29]. A q-dimensional simplex—the generalization of an edge—is a tuple of $q+1$ distinct nodes $(v_1, \ldots, v_{q+1})$; we say that this simplex is incident on $v_1, \ldots, v_{q+1}$. All simplices not included in a larger simplex are called the *facets* of the complex, whereas

*jean-gabriel.young.1@ulaval.ca

†alice.patania@isi.it

The sampling space would not be too constrained if these non-sequence-preserving bipartite graphs were rare. Sampling would then be easy. Unfortunately, it is straightforward to show that non-sequence-preserving graphs are far more common than sequence-preserving ones, by adapting the calculations of Ref. [31]. We find that the fraction ϕ of bipartite graphs with degrees $(\boldsymbol{d},\boldsymbol{s})$ not featuring parallel edges rapidly tends to

$$\phi = e^{-\frac{1}{2}(\langle d^2\rangle/\langle d\rangle - 1)(\langle s^2\rangle/\langle s\rangle - 1)}, \tag{2}$$

where $\langle x^k\rangle$ is the kth moment of the sequence $\boldsymbol{x}$, and where it is assumed that the elements of $\boldsymbol{d}$ and $\boldsymbol{s}$ do not grow with n (i.e., B is sparse). Thus, based on the presence of multiedges alone, there is a stringent upper bound on the fraction of bipartite graphs that are actually in the support of the SCM. An even smaller fraction remains after the bipartite graph with included neighborhood are removed.

B. Markov chain Monte Carlo method

To sample from the SCM, then, one needs to sample uniformly from a very constrained space, i.e., that of all sequence-preserving bipartite graphs with joint degree sequence $(\boldsymbol{d},\boldsymbol{s})$. Previously proposed approaches such as rejection sampling do not work well [22], because natural proposal distributions (e.g., stub matching) give an appreciable weight to non-sequence-preserving bipartite graphs [see Eq. (2)]. Thus, we turn to the Markov chain Monte Carlo (MCMC) sampling strategy [32], which has been used with great success for the CM [25,30]. The general idea is to construct a random chain of sequence-preserving bipartite graphs $B_0,\ldots,B_T$, to sample from this chain at regular intervals, and to treat the samples as if they had been drawn identically and independently from the ensemble. The algorithm will be correct if the chain is ergodic (time averages equal ensemble averages) and uniform (all nonisomorphic B are represented equally). These properties are determined by the allowed transformations $B_t \to B_{t+1}$ and the resulting transition matrix $\boldsymbol{\pi}$, where π_{ij} is the probability that B_j follows B_i in the chain. If the move set *connects the space* and the chain is *aperiodic*, then the chain will be ergodic. If the transition matrix is *doubly stochastic* (all rows and columns sum to 1), then the chain will be uniform.

We claim that the following set of moves satisfies all three conditions. Consider L, a random variable on the support $\mathcal{L} = \{2,3,\ldots,L_{\max}\}$, where $L_{\max}$ is a parameter and the distribution $\mathbb{P}[L=\ell]$ is arbitrary but nonzero everywhere on $\mathcal{L}$ (for illustration purposes, we will use $\mathbb{P}[L=\ell;\lambda] \propto e^{-\lambda\ell}$). At each step of the chain, we pick L edges in B (uniformly at random). We cut these edges and randomly match the stubs stemming from facets to the stubs stemming from nodes. If this matching generates a sequence-preserving bipartite graph B', then we accept the move; otherwise we resample B. This set of moves is similar to the double-edge swap commonly used in graph MCMC [25]. The only difference is the variable number of rewired edges, added to help the sampler better navigate the constrained support [30]. Much like its graphical counterpart, the resulting MCMC algorithm is efficient since drawing L edges and checking for resampling can be done in polynomial times.

The chain is aperiodic because the above set of moves yields a doubly stochastic transition matrix for any distribution $\mathbb{P}$: The total number of possible transitions at each configuration is a constant independent from the configuration considered (resampling guarantees this) [25]. The chain is also aperiodic, because there exists orbits of period 1 (resampling steps) and 2 (all moves are reversible) for any nontrivial $(\boldsymbol{d},\boldsymbol{s})$.

This leaves open the question of whether the support of the SCM is connected by the set of moves or not. We argue that it is, for all $L_{\max} \geqslant L^*_{\max}$, where $L^*_{\max}$ is bounded by

$$L^*_{\max} \leqslant 2\max \boldsymbol{s}. \tag{3}$$

To prove this, one would have to show that given two sequence-preserving bipartite graphs B_1 and B_2, it is always possible to find a B_3 such that $|\Delta_+[K(B_1),K(B_2)]| \geqslant |\Delta_+[K(B_1),K(B_3)]|$, where Δ_+ is the set of facets in $K(B_2)$ that are not in $K(B_1)$, and Δ_- is the set of facets in $K(B_1)$ that are not in $K(B_2)$ [$K(B)$ is the simplicial complex associated to the graph B]. Although a general proof remains elusive, we propose the following nonrigorous argument, valid for sparse simplicial complexes (simplicial complexes with bounded $\max \boldsymbol{d}$ and $\max \boldsymbol{s}$ in the limit $n \to \infty$).

To construct B_3, we first select a facet σ in Δ_+ (incident on the set of nodes Σ in B_2). The conservation of sizes and degrees guarantees that there exists a facet $\tau \in \Delta_-$ of the same size. The idea is then to start from B_1, cut all edges attached to τ and one edge from every node in Σ, match the stubs of σ to those of $v \in \Sigma$, and finally match the remaining orphaned stubs. This algorithm ensures that B_3 is closer to B_2 than B_1 was, because it removes facets from $\Delta_\pm$ (and does not add new facets either: Each $v \in \Sigma$ has at least on facet in Δ_- by the conservation of degrees). In general, it is not guaranteed that the last step can be carried out without creating included faces. However, in sparse simplicial complexes, σ is well separated from τ for almost all (σ,τ), since B_1 is locally treelike [28]. In such cases, no included faces are created at the last step, and the above algorithm can be carried through for some (σ,τ), generating B_3. Because this scheme involves at most $L^*_{\max} = 2\max \boldsymbol{s}$ rewired edges (when $|\tau| = |\sigma| = \max \boldsymbol{s}$), we obtain the bound of Eq. (3) for infinite sparse SCM. In practice, $L_{\max} = 2$ seems to always connect the space (we found no counterexamples), and sampling is more efficient when $L_{\max} \gg 2$ (see Fig. 3)—the value of $L^*_{\max}$ is more of theoretical than practical interest.

III. NULL MODEL

We put our efficient MCMC algorithm to the test, by verifying the statistical significance of the structural patterns found in three relational datasets that can be represented as simplicial complexes (see caption of Fig. 4 for details).

Since an instance of the SCM is provided in each case (the real system), we use it as the initial condition for each independent run of the sampling algorithm. Ergodicity implies that the state of the sampler will be uncorrelated with the initial configuration after a sufficiently long burn-in period—the choice of initial condition is ultimately irrelevant. Extrapolating from the results of Fig. 3, we opt for the proposal distribution $\mathbb{P}[L=\ell] = e^{\lambda\ell}/Z$ with $\lambda = 1$ and $L_{\max}$ set to 10% of $m = \sum d_i = \sum s_i$. Nonrigorous arguments from expander graph theory suggest $t_f = O(m\log m)$ as a good—if overzealous—choice of sampling interval [36].

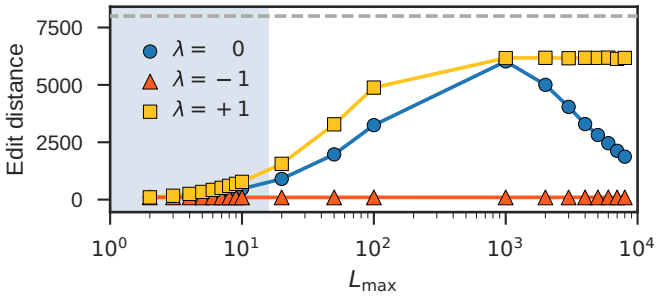

FIG. 3. Effect of the parametrization of the proposal distribution $\mathbb{P}$ on the mixing time, as quantified by the edit distances of the graphical representation of the samples. We investigate the family of distributions $\mathbb{P}[L = \ell; \lambda] = e^{\lambda \ell}/Z$, and use the regular SCM of $f = 1\,000$ facets of size $s = 8$, and $n = 2\,000$ nodes of degrees $d = 4$. Pairs of samples are separated by 100 proposed MCMC moves and are obtained from a unique initial configuration found via rejection sampling. The shaded region lies below the upper bound on $L^*_{\max}$ of Eq. (3). $\lambda = 1$ balances high-rejection probability but efficient moves with safe but inefficient moves, yielding the best overall performance for all $L_{\max}$. In practice, we have found that medium values of $L_{\max}$ are better, because checking for resampling is of complexity $O(L_{\max}\langle d \rangle)$, which translates into slower effective mixing time when $L_{\max} \gg 1$.

Significance results only make sense if they rely on a null model that embodies a natural null hypothesis for the problem at hand [25]. For example, the regular CM and its correlated variants usefully show that the network projection of datasets with high-order interactions are abnormally clustered [37]. Therefore, we use the sampler to investigate the distribution of a mesoscopic property only accessible when the datasets are encoded as simplicial complexes: The *shape* of the datasets, as captured by their homology, i.e., the pattern of holes, cavities, and higher dimensional voids [29]. The homology can be summarized by a series of Betti numbers $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots)$, where β_k counts the number of structural holes bounded by k-dimensional simplices. For example, β_0 counts the number of connected component, β_1 the number of homological cycles in K, β_2 the number of holes enclosed by facets of sizes 2, etc. Since every instance of the SCM has the same fixed local structure but is otherwise maximally random, we expect significant differences between the Betti number $\boldsymbol{\beta}$ of an organized simplicial complex and the bulk of the distribution of $\boldsymbol{\beta}$ in the corresponding randomized ensembles.

We show in Fig. 4 the distribution of β_0 and β_1 for the SCM associated to the real systems. Looking first at β_0, we find that the structure of the pollinator dataset is essentially random [Fig. 4(a)]. That is, the overwhelming majority of simplicial complexes with the same sequences have similar β_0. In contrast, the β_0 of the disease genome regulation (hereafter *diseasome*) and crime complexes are highly significant [Figs. 4(b)–4(c)]: A random instance of the SCM has fewer (diseasome) or more (crime) components than the real system with high probability. In one case (crime), the difference is a statistical signature of how the dataset was gathered, namely by looking up the ties of suspects, victims, and witnesses already in the dataset, recursively [35]. Because this process creates much larger connected components than random sampling, the resulting β_0 is far from the ensemble average—an effect that we expect to find in any dataset

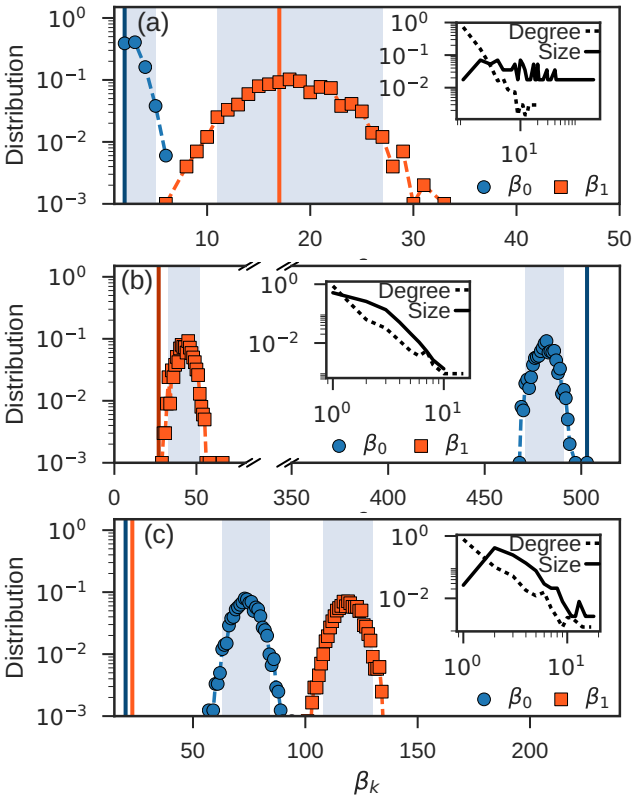

FIG. 4. Significance of the Betti numbers of real systems. The datasets are bipartite networks, which we convert to simplicial complexes (we prune included faces). They map the relationships between (a) flower-visiting insects (nodes, $n = 679$) and plants (facets $f = 57$) in Kyoto [33], (b) human disease (nodes $n = 1100$) and genes (facets $f = 752$) linked by known disorder-gene associations [34], and (c) crimes (nodes, $n = 829$) and suspects, victims, and witnesses (facets, $f = 378$) in St. Louis [35]. The Betti numbers of these real systems appear as solid vertical lines and are equal to (a) $\beta_0 = 2$, $\beta_1 = 17$ (b) $\beta_0 = 503$, $\beta_1 = 27$, and (c) $\beta_0 = 20$, $\beta_1 = 23$. We show the distributions of Betti numbers for the equivalent SCM with solid symbols (computed from 1000 instances of the model). The shaded regions contain 95% of the samples. The parameters of the SCM—extracted from real systems—are shown in insets.

constructed using a similar methodology. In the other case (diseasome), the real system has *more* components than one would typically expect from the local information alone. The construction procedure does not explain this disparity [34], meaning that the system must self-organize in a fragmented way, likely for biological or evolutionary reasons.

Turning to β_1 we again find that the structure of the pollinator dataset is typical and that the same cannot be said of the diseasome and crime datasets. Both simplicial complexes have significantly fewer cycles than expected; i.e., given a cycle, it is more likely to be filled by a simplex in the real system than in the randomized one, suggesting that some form of high-order triadic closure is at play [10]. The difference is, however, much more pronounced in the crime dataset; this could be due to the fact that it describes a social system, whose structure tend to be heavily driven by triadic closure [38] (and potential high order analogs).

Finally, taking both distributions into account, we conclude that the shape of the pollinator dataset is completely determined by its local structure, while large-scale organizational principles influence the structure of the other datasets. This leads us to two final observations: One, care must be exerted in drawing conclusions about the shape of complex datasets—from the homology point of view there is nothing of note in the structure of the pollinator dataset. Two, some datasets—here the crime and diseasome datasets—are decidedly *not* random. This raises the question of just how much information must models account for before they can capture such atypical Betti numbers. Would, for example, adding limited correlations among degrees be sufficient to capture the shape of most real datasets? Or do we need to embrace growth models, with their sophisticated rules and clustered local structure [8,23,24]?

IV. PERSPECTIVES

As it stands, the SCM already establishes the analysis of simplicial complexes on firmer statistical ground. The next step will be to clarify a number of important open questions, e.g., what is the true value of $L^*_{\max}$ for arbitrary simplicial complexes and what is optimal choice of proposal distribution $\mathbb{P}$ (cf. Fig. 3).

Beyond these obvious questions, the connection between the SCM and the simple CM lead us to a series of natural problems not addressed in this paper. These include the problem of the *simpliciality* of arbitrary pairs of sequences (i.e., is there a simplicial complex which realize a pair of sequences?) [22], related to the problem of constructing initial conditions for the MCMC sampler, when no real system is available. We believe that the solution to such problems will require new insights, as the no-inclusion constraints appear to be a major obstacle to the application of classical methods developed for the analogous *graphicality* problem [39,40].

In closing, we stress that all the above questions and challenges are of technical nature; the model and sampler can already be applied to practical problems [32]. This could lead to improvements in persistent homology (e.g., statistically sound filtrations of weighted complexes) or a formulation of community detection of simplicial complexes (via modularity [41]) and could provide a new glimpse into the emergence of homology and higher order structural properties in real complex systems.

ACKNOWLEDGMENTS

We thank L. Hébert-Dufresne and G. Bianconi for helpful discussions and comments. The authors acknowledge the support of the ADnD project by Compagnia San Paolo (A.P., G.P.), the Fonds de recherche du Québec-Nature et technologies (J.G.Y.), the Complex Systems Langrange Lab (F.V.), and the YRNCS Bridge Grant (A.P., J.G.Y.). A.P. is grateful for the hospitality of L. J. Dubé at Université Laval, where parts of the research work was conducted.

A.P. and J.G.Y. contributed equally to this work.

[1] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, Rev. Mod. Phys. **87**, 925 (2015).
[2] Y.-Y. Liu and A.-L. Barabási, Rev. Mod. Phys. **88**, 035006 (2016).
[3] M. A. Porter, J.-P. Onnela, and P. J. Mucha, Notices AMS **56**, 1082 (2009).
[4] M. E. J. Newman, Nat. Phys. **8**, 25 (2012).
[5] Y. Dabaghian, F. Mémoli, L. Frank, and G. Carlsson, PLoS Comput. Biol. **8**, e1002581 (2012).
[6] C. Giusti, R. Ghrist, and D. S. Bassett, J. Comput. Neurosci. **41**, 1 (2016).
[7] K. Xia and G.-W. Wei, Int. J. Numer. Methods Biomed. Eng. **30**, 814 (2014).
[8] L. Hébert-Dufresne, E. Laurence, A. Allard, J.-G. Young, and L. J. Dubé, Phys. Rev. E **92**, 062809 (2015).
[9] B. Stolz, H. Harrington, and M. A. Porter, arXiv:1610.00752 (unpublished).
[10] K. Zuev, O. Eisenberg, and D. Krioukov, J. Phys. A **48**, 465002 (2015).
[11] S. P. Ellis and A. Klein, Homology, Homotopy Appl. **16**, 245 (2014).
[12] C. Curto and V. Itskov, PLoS Comput. Biol. **4**, e1000205 (2008).
[13] D. Horak, S. Maletić, and M. Rajković, J. Stat. Mech. Theor. Exp. (2009) P03034.
[14] A. Patania, F. Vaccarino, and G. Petri, EPJ Data Sci. **6**, 7 (2017).
[15] G. Petri, M. Scolamiero, I. Donato, and F. Vaccarino, PloS One **8**, e66506 (2013).
[16] A. Sizemore, C. Giusti, and D. S. Bassett, J. Complex Netw. **5**, 245 (2017).
[17] J. M. Chan, G. Carlsson, and R. Rabadan, PNAS **110**, 18566 (2013).
[18] G. Petri, P. Expert, F. Turkheimer, R. Carhart-Harris, D. Nutt, P. Hellyer, and F. Vaccarino, J. R. Soc. Interface **11**, 20140873 (2014).
[19] Y. Hiraoka, T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, and Y. Nishiura, PNAS **113**, 7035 (2016).
[20] M. Kahle, AMS Contemp. Math. **620**, 201 (2014).
[21] A. Costa and M. Farber, in *Configuration Spaces* (Springer, Berlin, 2016), pp. 129–153.
[22] O. T. Courtney and G. Bianconi, Phys. Rev. E **93**, 062311 (2016).
[23] G. Bianconi and C. Rahmede, Phys. Rev. E **93**, 032315 (2016).
[24] Z. Wu, G. Menichetti, C. Rahmede, and G. Bianconi, Sci. Rep. **5**, 10073 (2015).
[25] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, arXiv:1608.00607 (unpublished).
[26] C. Orsini, M. M. Dankulov, A. Jamakovic, P. Mahadevan, P. Colomer-de Simón, A. Vahdat, K. E. Bassler, Z. Toroczkai, M. Boguñá, G. Caldarelli *et al.*, Nat. Commun. **6**, 8627 (2015).
[27] M. Molloy and B. A. Reed, Rand. Struct. Alg. **6**, 161 (1995).
[28] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Phys. Rev. E **64**, 026118 (2001).
[29] A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, UK, 2000).
[30] I. Miklós, P. L. Erdős, and L. Soukup, Electron. J. Combin. **20**, P16 (2013).
[31] E. A. Bender and E. R. Canfield, J. Combin. Theo. A **24**, 296 (1978).

[32] We provide a reference C++ implementation of the sampler as well as tutorials at https://www.github.com/jg-you/scm.
[33] M. Kato, T. Kakutani, T. Inoue, and T. Itino, Contr. Biol. Lab. Kyoto Univ. **27**, 309 (1990).
[34] K.-I. Goh, M. E. Cusick, D. Valle, B. Childs, M. Vidal, and A.-L. Barabási, PNAS **104**, 8685 (2007).
[35] S. Decker, C. W. Kohfeld, R. Rosenfeld, and J. Sprague, *St. Louis Homicide Project: Local Responses to a National Problem* (University of Missouri, St. Louis, 1991).
[36] We represent the support of the SCM as a graph $\mathcal{G}(L_{\max}) = (V, E)$. If $\mathcal{G}$ is an expander, then the sampler yields uncorrelated configuration with high probability after $t_f = O(\log |V|)$ steps; the suggested t_f follows from the loose upper bound $|V| \leqslant m!$. A proof that $\mathcal{G}(L_{\max})$ is in fact an expander will depend on $(\boldsymbol{d}, \boldsymbol{s}, L_{\max})$; however, we note that $\mathcal{G}(L_{\max})$ shares two important properties with all expanders for sufficiently large $L_{\max}$: It is connected and not bipartite.
[37] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).
[38] C. A. Hidalgo, Applied Network Science **1**, 6 (2016).
[39] V. Havel, Casopis Pest. Mat. **80**, 477 (1955).
[40] S. L. Hakimi, J. Soc. Ind. Appl. Math. **10**, 496 (1962).
[41] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).