18th International Conference on Information Fusion
Washington, DC - July 6-9, 2015

Hidden network reconstruction from information diffusion

Forrest W. Crawford
Department of Biostatistics
Yale School of Public Health
New Haven, Connecticut 06510
http://crawford.research.yale.edu

***Abstract*—Learning about the structure of hidden or covert networks is a major challenge in epidemiology, sociology, and intelligence analysis. Vertices in hidden networks usually cannot be enumerated or sampled in a systematic way; they can only be revealed by tracing links emanating from already-observed vertices. Observers sometimes cannot follow links directly, and instead must rely on passive observation of a dynamic process to reveal vertices and edges. This paper outlines a framework for estimating network structures from partial observation of information diffusion through the network. Diffusion is modeled by a continuous-time Markov epidemic model. Edges are revealed by transmission events and new vertices are uncovered when information is transmitted to them. The approach is a generalization of tools developed to reconstruct drug-user networks from respondent-driven sampling studies in epidemiology. The likelihood of the diffusion process can be interpreted as an exponential random graph model. A Bayesian method for probabilistic reconstruction of the transmission-induced subgraph is described.**

Keywords: Covert network, diffusion, exponential random graph model, epidemic model, Markov process, network reconstruction

I. Introduction

Many social, organizational, and operational networks are obscured from view and cannot be studied by comprehensive census. When vertices can be randomly sampled, it is sometimes possible to estimate global or local graph properties from a sampling-induced subgraph [1]–[8]. Unfortunately random sampling of vertices or edges in a hidden network is often impossible because no appropriate sampling frame is available. When some or all vertices of a hidden network are observed, explicit reconstruction often takes the form of link prediction [9]–[14]. When most of the network is observed, but some vertices are hidden or covert, it may be possible to detect missing vertices [15]. When vertices or edges can be observed in more than one sample, it is often possible to combine information across samples to reconstruct subgraphs or estimate global graph properties [16]–[18]. However, estimation techniques that are effective under random sampling do not necessarily perform well under other observation scenarios [19]. In particular, estimates of global graph properties from observed vertices – such as the degree distribution – may be strongly affected by the sampling procedure [20], [21].

Hidden or covert networks can typically only be studied by tracing links from one vertex to another. Link-tracing survey techniques have gained wide use in epidemiology and sociology for studies of hidden or hard-to-reach populations [22], [23]. Often social stigma serves to obscure members of hidden populations; sometimes the fear of legal repercussions keeps individuals hidden. Respondent-driven sampling (RDS) [23] is a link-tracing procedure that has found wide use in epidemiology, sociology, and public health research on drug users. RDS has also been proposed to study domestic extremism and counterinsurgency [24]–[26].

Link-tracing studies aim to discover new nodes and the connections between them. When the link-tracing process constitutes a sampling design, it is possible to estimate global properties of the network and characteristics of the population of vertices [27]. There is evidence that data from link-tracing studies can have very different properties from data obtained by random sampling of vertices [28]–[30]. In link-tracing studies, not every link between sampled nodes is observed; usually only links that are traced can be observed, and it is unclear whether analysts can hope to estimate properties of the local or global networks from networks observed in this way.

Worse, the mechanism that reveals links is often not under the control of the observer. For example, in RDS subjects “recruit” other subjects to whom they are connected in the target population social network. As another example, an intelligence analyst might intercept a message diffusing through a covert network. In this idealization, vertices and edges are revealed to the observer only when an action – such as communication or transmission of information – happens across the edge. If no transmission happens across a given edge, the edge is not revealed to the observer. Likewise, if no transmission reaches a given vertex, that vertex is not revealed to the observer. Despite these limitations, passive observation of deterministic or stochastic communication processes on links has been effectively used to provide insight into the structure of hidden or partially obscured networks [31]–[34]. The key insight in this work is that the path and dynamics of a process on a hidden network can reveal properties of the network itself.

This paper outlines a general strategy for probabilistic reconstruction of the edges in a hidden network from observation of an information diffusion process on that network. Information diffusion is modeled by a continuous-time Markov susceptible-infected model [35]. This work is disctinct from

978-0-9824-4386-6/15/$31.00 ©2015 IEEE

"diffusion" methods for distributed learning over networks [36]–[38]. The technique is a generalization of tools developed for epidemiological research on networks of drug users from data obtained by RDS link tracing [39]. First, a class of Markov transmission processes is defined, along with the data observed from the process. Vertices are observed when information is transmitted to them; edges are visible only when a transmission event takes place across them. The notion of "transmission" is general: it can refer to any one-way communication process that changes the state of the individual who receives the transmission, and in which the receiver can also transmit the information to its network neighbors. Passive observation of this type of communication processes can reveal important properties of network structures. A Bayesian method for probabilistic reconstruction of the transmission-induced subgraph is derived.

II. Markov diffusion processes on networks

A. Preliminaries

Suppose we wish to learn about the structure of an undirected graph $G = (V, E)$, where the vertex set V has finite size $|V| = N < \infty$ and the edge set E contains no self-loops or parallel edges. A vertex's degree is the number of edges incident to it that connect to other vertices in the hidden network G. The terms "graph" and "network", "vertex" and "node", and "edge" and "link" are used interchangeably. The graph could represent a social network, an organizational structure, or relationships between any entities of interest. A stochastic model of one-way information diffusion on the edges of G is constructed in a manner analogous to the susceptible-infected model of infectious disease epidemiology [35]. The term "diffusion" refers to the spread or transmission of a state, message, or object along the edges of G.

Suppose that each vertex $i \in V$ has a property or state $X_i(t)$ which is a function of time $t > 0$. In this paper, it is assumed for simplicity that $X_i(t) \in \{0, 1\}$, but this restriction could be relaxed. A vertex in state 1 at time t has already received the message, and can transmit it; a vertex in state 0 has not yet received it. Assume that at time $t = 0$, the set of vertices M with $X_i(0) = 1$ for $i \in M$, is known. We refer to members of M as "seeds".

Definition 1 (Susceptible vertices and edges)**.** *A vertex $j \in V$ is susceptible to transmission at time t if $X_j(t) = 0$ and there exists least one $i \in V$ such that $X_i(t) = 1$ and $\{i, j\} \in E$. An edge $\{i, j\} \in E$ is susceptible at time t if $X_i(t) = 1$ and $X_j(t) = 0$ or $X_i(t) = 0$ and $X_j(t) = 1$.*

The time to transmission along a susceptible edge is assumed to follow a common probability distribution. Consider two distinct vertices $i \in V$ and $j \in V$ with $\{i, j\} \in E$. At time $t = 0$, assume that $X_i(0) = 1$ and $X_j(0) = 0$. Transmission happens independently across the edge connecting i and j at a random time

$$T_{ij} \sim \text{Exponential}(\lambda), \qquad (1)$$

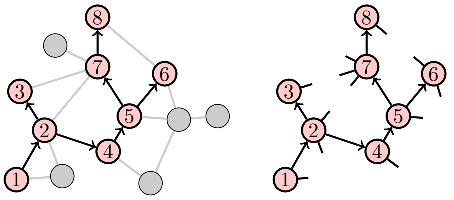

Fig. 1. Transmission path on the true graph and observed transmission path. At left is the transmission path G_T overlaid on the network G. At left is the observed transmission path with pendant edges implied by the observed degrees. In particular, the analyst does not observe the full network G, nor the transmission-induced subgraph G_S.

where $\lambda > 0$ is the rate of transmission across a single susceptible edge.

B. Observation

Suppose the vertex $i \in V$ becomes known to an observer at time $t_i = \text{argmin}_t\{t : X_i(t) = 1\}$ when it first receives a transmission. The degree d_i of i in G is fully observed. By assumption, the observer does not have direct access to any vertices or edges in G; instead, the path of a stochastic transmission process on the edges of G is revealed over time.

Definition 2 (Transmission graph)**.** *The directed transmission graph is $G_T = (V_T, E_T)$, where $V_T \subset V$ is the set of n known vertices and a directed edge $(i, j) \in E_T$ indicates that i transmitted a message to j.*

It is assumed that vertices cannot receive a transmission more than once, so G_T is acyclic. This assumption can be relaxed with some increase in notation and computation. Furthermore, G_T need not be connected if the set of seeds has $|M| > 1$.

While the directed transmission graph G_T is fully observed, the subgraph of observed vertices is not visible, since an edge between vertices in V_T is not visible unless a transmission event took place across that edge.

Definition 3 (Transmission-induced subgraph)**.** *The transmission-induced subgraph is an undirected graph $G_S = (V_S, E_S)$, where $V_S = V_T$ consists of n sampled vertices, and $\{i, j\} \in E_S$ if and only if $i \in V_S$, $j \in V_S$, and $\{i, j\} \in E$.*

From this definition, it is evident that the recruitment graph G_T is a directed subgraph of G_S.

Definition 4 (Transmissibility matrix)**.** *Let $\mathbf{T}$ be a $n \times n$ matrix whose element $\mathbf{T}_{ij}$ is 1 if vertex i can transmit the information just before the time of the jth transmission event, and zero otherwise. The rows and columns of $\mathbf{T}$ are ordered by the time at which each subject is observed.*

Let $\mathbf{d}$ be the time-ordered $n \times 1$ vector of subjects' degrees in the order they are observed and let $\mathbf{t} = (t_1, \ldots, t_n)$ be the $n \times 1$ vector of transmission times, where $t_1 < \cdots < t_n$. The observed data from the transmission process consists of $\mathbf{Y} = (G_T, \mathbf{d}, \mathbf{t}, \mathbf{T})$.

Figure 1 illustrates the observed data and their relationship to the unobserved population graph G. Since the transmission graph G_T does not contain any edges along which a transmission event did not take place, the transmission-induced subgraph G_S is not fully observed. However, observation of G_T and $\mathbf{d}$ places constraints on the topology of G_S.

C. Likelihood

To derive the likelihood of the observed data on a hidden network, it is necessary to formalize the class of reconstructed networks for which the likelihood makes sense.

Definition 5 (Compatibility)**.** *A subgraph $\widehat{G}_S = (V_S, \widehat{E}_S)$ is compatible with the transmission graph G_T if*

1) *for each $(i,j) \in E_T$, $\{i,j\} \in \widehat{E}_S$;*
2) *the degree of $i \in V_S$ in G_S does not exceed the observed total degree d_i in G.*

Intuitively, an estimated subgraph $\widehat{G}_S$ is compatible with G_T if G_T is a (directed) subgraph of G_S and the degree of each vertex in G_S is less than or equal to its degree in G. Let $\mathbf{w} = (0, t_1, t_2 - t_1, \ldots, t_n - t_{n-1})$ be the $n \times 1$ vector of inter-transmission waiting times and let $\mathbf{A}$ be the $n \times n$ adjacency matrix of $\widehat{G}_S$, with the rows and columns representing vertices in the order they were observed. Let $\mathbf{u}$ be an $n \times 1$ vector whose ith element is the number of edges connecting i to unobserved vertices in G, so $\mathbf{u}_i = d_i - \sum_{j=1}^{n} \mathbf{A}_{ij}$. Figure 2 shows the matrices used to compute the likelihood. The joint likelihood of G_T and $\mathbf{w}$ can be expressed in a computationally convenient form without explicitly enumerating susceptible edges. The likelihood is given by

$$L(G_T, \mathbf{w} | G_S, \mathbf{d}, \lambda) = \lambda^{n-|M|} \exp[-\lambda \mathbf{s}'\mathbf{w}] \qquad (2)$$

where

$$\mathbf{s} = \text{lowerTri}(\mathbf{AT})'\mathbb{1} + \mathbf{T}'\mathbf{u} \qquad (3)$$

Crawford (2015) gives a proof [39]. The statistic $\mathbf{s}$ is a $n \times 1$ vector whose ith element is the number of susceptible edges in G_S just before the ith transmission event.

III. Network reconstruction

A. Exponential random graph models

The likelihood (2) can be interpreted as a function of the adjacency matrix $\mathbf{A}$, with $\mathbf{w}$ and λ held fixed. We can rewrite (2) as

$$\Pr(\mathbf{A}) = \frac{\exp[\mathbf{s}(\mathbf{A})'\boldsymbol{\theta}]}{\kappa(\boldsymbol{\theta})} \qquad (4)$$

where $\mathbf{s}(\mathbf{A})$ is a vector-valued function of $\mathbf{A}$ given by (3), $\boldsymbol{\theta} = -\lambda \mathbf{w}$, and $\kappa(\boldsymbol{\theta})$ is a normalizing constant that does not depend on $\mathbf{A}$. The interpretation (4) reveals that the likelihood of the observed data constitutes an exponential random graph model (ERGM) for the unobserved portion of the transmission-induced subgraph G_S [40], [41]. ERGMs have several desirable computational properties: it is straightforward to simulate realizations from (4) via Gibbs sampling or Metropolis-Hastings steps; to compute the ratio of probabilities of two estimated graphs, only a change statistic needs to be calculated [42].

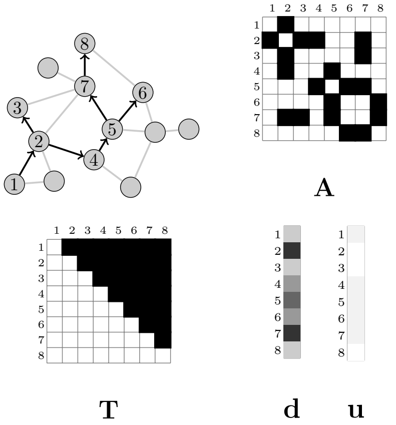

Fig. 2. Matrices used to compute the likelihood (2).

B. Generating a compatible estimate

Given a compatible graph $G_S = (V_S, E_S)$ with corresponding vector $\mathbf{u}$, another compatible estimate can be obtained by randomly choosing two vertices i and j with $t_i < t_j$ such that no transmission event took place from i to j, that is, $(i,j) \notin E_T$. If $\{i,j\} \notin E_S$, $\mathbf{u}_i > 0$ and $\mathbf{u}_j > 0$, then a new edge $\{i,j\}$ is proposed. If $\{i,j\} \in E_S$, then it is proposed to remove the edge $\{i,j\}$ from E_S. The resulting proposal graph is identical to G_S except that the edge $\{i,j\}$ has either been added or removed. Furthermore, since this procedure does not change transmission edges in G_T, the proposal graph is compatible with the observed data, by Definition 5.

The number of compatible subgraphs G_S^* that can be produced from G_S by this procedure is

$$\begin{aligned}\text{nchanges}(G_S) = \sum_{i<j} & \mathbb{1}\{\{i,j\} \notin E_S, \mathbf{u}_i > 0, \mathbf{u}_j > 0\} \\ & + \mathbb{1}\{\{i,j\} \in E_S, (i,j) \notin E_R\}.\end{aligned} \qquad (5)$$

The probability of producing any particular compatible subgraph G_S^* is

$$\Pr(G_S^* | G_S) = \frac{1}{\text{nchanges}(G_S)}. \qquad (6)$$

This probability will be useful in forming the Metropolis-Hastings ratio for G_S^* below.

C. Computing the likelihood ratio

Suppose first that G_S has no edge between i and j, $\{i,j\} \notin E_S$. For a proposal $G_S^+ = (V_S, E_S^+)$ identical to G_S except that $\{i,j\} \in E_S^+$, the likelihood ratio is

$$\frac{L(G_T, \mathbf{w} | G_S^+, \mathbf{d}, \lambda)}{L(G_T, \mathbf{w} | G_S, \mathbf{d}, \lambda)} = e^{2\lambda(t_n - t_j)}. \qquad (7)$$

Now suppose G_S has and edge between i and j, $\{i,j\} \in E_S$ with $\{i,j\} \notin E_T$. For a proposal $G_S^- = (V_S, E_S^-)$ identical to G_S except that $\{i,j\} \notin E_S^-$, the likelihood ratio is

$$\frac{L(G_T, \mathbf{w} | G_S^-, \mathbf{d}, \lambda)}{L(G_T, \mathbf{w} | G_S, \mathbf{d}, \lambda)} = e^{-2\lambda(t_n - t_j)}. \qquad (8)$$

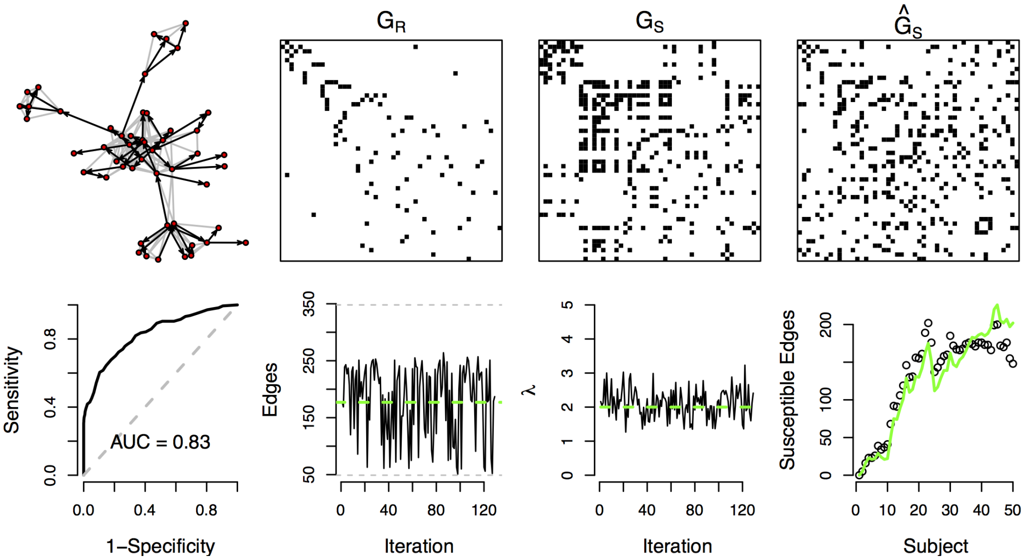

Fig. 3. Example Bayesian reconstruction of transmission-induced subgraph G_S from simulated transmission process. The network data are from a study of social, sexual, and drug use links between individuals in Colorado Springs, CO, USA from 1988-1990. An information diffusion process is simulated through the network with $|M| = 1$ seed, $n = 50$ total observations, and transmission rate $\lambda = 2$. The prior parameters for λ are $\alpha = 1$, $\beta = 0.57$, and the sparsity penalty is $\gamma = 1.77$. The top row shows the true transmission-induced subgraph G_S, with transmission edges shown in black and unobserved edges in gray. The adjacency matrices of G_T, G_S, and a random draw $\widehat{G}_S$ from the posterior distribution are also shown. The receiver-operating characteristic (ROC) curve is shown with the area under the curve (AUC) indicating good overall reconstruction accuracy. The posterior traces of the number of edges, λ, and susceptible edges are shown. True values are given in green.

These expression are simple and do not require the any matrix computation implied by (2) and (3). Furthermore, the differences $t_n - t_j$ can be computed in advance and stored for repeated use.

D. Prior for G_S

While the likelihood (2) helps determine which edges belong in G_S, point estimates or Bayesian posterior estimates can contain more edges than the true G_S. It is clear from (7) and (8) that addition of a new edge $\{i, j\}$ in G_S is favored over sending both pendant edges to unobserved vertices not in G_S. A prior distribution for G_S is therefore helpful to ensure the desired sparsity. A convenient class of priors is $\Pr(G_S) \propto \exp[-\gamma|E_S|]$, where $\gamma > 0$ is chosen to penalize dense graphs. Given G_T and $\mathbf{d}$, we have the inequalities

$$n - |M| \le |E_S| \le \frac{1}{2}\sum_{i=1}^{n} \mathbf{d}_i, \tag{9}$$

and these bounds can be used to specify γ. The lower bound is sharp (it is the number of edges in the transmission subgraph G_T), while the upper bound could be tightened with specific knowledge of $\mathbf{d}$. Let $p = \left(n - |M| + \frac{1}{2}\sum_i \mathbf{d}_i\right)/2\binom{n}{2}$ be a crude estimate of the density of G_S. Then letting $\gamma = -\log[p/(1-p)]$ gives a convenient sparsity penalty without imposing undue assumptions on the topology of G_S.

E. Algorithm

Using the results given above, it is possible to reconstruct the transmission-induced subgraph G_S and observed vertices' connections to unobserved vertices in G with reasonable accuracy. By alternately drawing from the conditional distributions of G_S and λ, a Gibbs sampling algorithm is derived. First suppose λ is held fixed, and G_S is a compatible subgraph estimate. From this estimate, a new compatible subgraph G_S^* is formed using the procedure outlined in Section III-B. Then we compute the Metropolis-Hastings ratio

$$\frac{L(G_T, \mathbf{w}|G_S^*, \mathbf{d}, \lambda)}{L(G_T, \mathbf{w}|G_S, \mathbf{d}, \lambda)} \cdot \frac{\Pr(G_S^*)}{\Pr(G_S)} \cdot \frac{\Pr(G_S|G_S^*)}{\Pr(G_S^*|G_S)} \tag{10}$$

and accept G_S^* if this ratio is greater than 1. Otherwise, we accept G_S^* with probability equal to this ratio.

Next, suppose G_S is fixed and we wish to sample λ. We employ a Gamma prior for λ with $\pi(\lambda) \propto \lambda^{\alpha-1}e^{-\beta\lambda}$ where $\alpha > 0$ and $\beta > 0$. Then $\pi(\lambda)$ is a conjugate prior for the likelihood (2), and the conditional posterior distribution of λ is $\text{Gamma}(\alpha + n - |M|, \beta + \mathbf{s}'\mathbf{w})$. Therefore, we can sample λ directly from its conditional posterior distribution.

IV. Results

Figure 3 shows an example of Bayesian reconstruction of G_S. The network $G = (V, E)$ is derived from a network study of social, sexual, and drug use links between individuals in Colorado Springs, CO, USA from 1988-1990 [43]–[45] with $|V| = 5492$ and $|E| = 43288$. An information diffusion process is simulated on G with $|M| = 1$ seeds, $n = 50$ observations, and $\lambda = 2$. The prior for λ is Gamma with $\alpha = 1$, $\beta = 0.57$ (giving prior mean $\mathbb{E}[\lambda] = 2$), and the sparsity penalty is $\gamma = 1.77$. The marginal edge-wise posterior distribution of G_S is used to assess reconstruction performance. The top row of Figure 3 shows the transmission-induced subgraph G_S with transmission edges in black and unobserved edges in gray; G_T, G_S, and a random draw $\widehat{G}_S$ from the posterior distribution are shown. The bottom row shows the receiver-operating characteristic (ROC) curve with area under the curve (AUC) 0.83, indicating good reconstruction accuracy. Next, the number of edges in the reconstructed estimates $\widehat{G}_S$, estimates of λ, and the number of susceptible edges in $\widehat{G}_S$ are shown, with true values given in green lines. Gray dashed lines in the trace of edge counts denote the minimum and maximum edge counts given G_T and $\mathbf{d}$ described in (9). A modified application of this approach to drug user networks is given by [39].

V. Conclusions

Mapping of covert networks is an important task in intelligence analysis and threat detection [46], [47]. Discovery of nodes and links can be challenging, especially when analysts must rely on passive observation for insight. This paper has shown that observation of a diffusion process can reveal topological properties of a hidden network. Two features of the proposed method yield desirable inferential properties. First, compatibility (in the sense of Definition 5) induces strong topological constraints on estimated subgraphs, but without additional insight, all such compatible subgraphs have the same probability. Second, the likelihood of a stochastic diffusion process can used to distinguish between compatible topologies, providing more weight to those that occur with higher probability under the diffusion model. Furthermore, it may be possible to extend the proposed tools for estimating the transmission-induced subgraph to estimation of features of the unobserved parts of the super-population graph G. When a model is specified for the global network, the sampled portion of the graph can sometimes be used to probabilistically impute the remaining part [48].

The stochastic model of information diffusion employed here is simple and parsimonious, and is based on widely used models of epidemic processes on graphs. More complicated models that incorporate loss of transmissibility (entailing changes to the structure of the transmissibility matrix $\mathbf{T}$), or preferential transmission between certain types of vertices are possible with little additional computational burden. However, assumptions required by more complicated models may not be justifiable when aspects of the diffusion process are not known with certainty, and a balance is necessary between realism and parsimony.

Acknowledgments

I thank Robert Heimer, Edward M. Kaplan, Jiacheng Wu, and Li Zeng for valuable comments and criticism. This work is supported by NIH/NCATS grant KL2 TR000140, NIMH grant P30MH062294, and startup funds from the Yale School of Public Health. I acknowledge the Yale University Biomedical High Performance Computing Center for computing support, funded by NIH grants RR19895 and RR029676-01.

References

[1] O. Frank, "Sampling and estimation in large social networks," *Social networks*, vol. 1, no. 1, pp. 91–101, 1978.

[2] ——, "Sampling and inference in a population graph," *International Statistical Review/Revue Internationale de Statistique*, pp. 33–41, 1980.

[3] J. Galaskiewicz, "Estimating point centrality using different network sampling techniques," *Social Networks*, vol. 13, no. 4, pp. 347–386, 1991.

[4] E. Costenbader and T. W. Valente, "The stability of centrality measures when networks are sampled," *Social networks*, vol. 25, no. 4, pp. 283–307, 2003.

[5] S. P. Borgatti, K. M. Carley, and D. Krackhardt, "On the robustness of centrality measures under conditions of imperfect data," *Social networks*, vol. 28, no. 2, pp. 124–136, 2006.

[6] G. Kossinets, "Effects of missing data in social networks," *Social networks*, vol. 28, no. 3, pp. 247–268, 2006.

[7] S. González-Bailón, N. Wang, A. Rivero, J. Borge-Holthoefer, and Y. Moreno, "Assessing the bias in samples of large online networks," *Social Networks*, vol. 38, pp. 16–27, 2014.

[8] R. Grannis, "Sampling effects in social network analysis," in *Encyclopedia of Social Network Analysis and Mining*, R. Alhajj and J. Rokne, Eds. Springer, New York, 2014.

[9] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American society for information science and technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[10] J. Leskovec, D. Huttenlocher, and J. Kleinberg, "Predicting positive and negative links in online social networks," in *Proceedings of the 19th International Conference on World Wide Web*. ACM, 2010, pp. 641–650.

[11] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[12] Y. F. Atchade, "Estimation of network structures from partially observed markov random fields," *arXiv preprint arXiv:1108.2835*, 2011.

[13] J. H. Koskinen, G. L. Robins, P. Wang, and P. E. Pattison, "Bayesian analysis for partially observed network data, missing ties, attributes and actors," *Social Networks*, vol. 35, no. 4, pp. 514–527, 2013.

[14] C. A. Bliss, C. M. Danforth, and P. S. Dodds, "Estimation of global network statistics from incomplete data," *PloS One*, vol. 9, no. 10, p. e108471, 2014.

[15] Y. Maeno, "Node discovery in a networked organization," in *Systems, Man and Cybernetics, 2009. SMC 2009. IEEE International Conference on*. IEEE, 2009, pp. 3522–3527.

[16] O. Frank and T. Snijders, "Estimating the size of hidden populations using snowball sampling," *Journal of Official Statistics*, vol. 10, pp. 53–53, 1994.

[17] J. A. Smith, "Macrostructure from microstructure generating whole systems from ego networks," *Sociological Methodology*, vol. 42, no. 1, pp. 155–205, 2012.

[18] B. Yan and S. Gregory, "Identifying communities and key vertices by reconstructing networks from samples," *PloS one*, vol. 8, no. 4, p. e61006, 2013.

[19] P. Ebbes, Z. Huang, and A. Rangaswamy, "Subgraph sampling methods for social networks: The good, the bad, and the ugly," *HEC Paris Research Paper No. MKG-2014-1027*, 2013.

[20] M. P. Stumpf, C. Wiuf, and R. M. May, "Subnets of scale-free networks are not scale-free: sampling properties of networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4221–4224, 2005.

[21] S. H. Lee, P.-J. Kim, and H. Jeong, "Statistical properties of sampled networks," *Physical Review E*, vol. 73, no. 1, p. 016102, 2006.

[22] L. A. Goodman, "Snowball sampling," *The Annals of Mathematical Statistics*, vol. 32, no. 1, pp. 148–170, 1961.

[23] D. D. Heckathorn, "Respondent-driven sampling: a new approach to the study of hidden populations," *Social Problems*, vol. 44, no. 2, pp. 174–199, 1997.

[24] R. Silva, J. Klingner, and S. Weikart, "Measuring lethal counterinsurgency violence in Amritsar District, India using a referral-based sampling technique," in *Joint Statistical Meetings*, 2010, pp. 552–580.

[25] G. Davies and S. Dawson, "A framework for estimating the number of extremists in Canada," 2014.

[26] P. Joosse, S. M. Bucerius, and S. K. Thompson, "Narratives and counternarratives: Somali-Canadians on recruitment as foreign fighters to Al-Shabaab," *British Journal of Criminology*, 2015.

[27] S. K. Thompson and O. Frank, "Model-based estimation with link-tracing sampling designs," *Survey Methodology*, vol. 26, no. 1, pp. 87–98, 2000.

[28] B. H. Erickson, "Some problems of inference from chain data," *Sociological Methodology*, vol. 10, no. 1, pp. 276–302, 1979.

[29] P. Biernacki and D. Waldorf, "Snowball sampling: problems and techniques of chain referral sampling," *Sociological Methods & Research*, vol. 10, no. 2, pp. 141–163, 1981.

[30] M. De Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher, "How does the data sampling strategy impact the discovery of information diffusion in social media?" *ICWSM*, vol. 10, pp. 34–41, 2010.

[31] M. A. Kramer, U. T. Eden, S. S. Cash, and E. D. Kolaczyk, "Network inference with confidence from multivariate time series," *Physical Review E*, vol. 79, no. 6, p. 061916, 2009.

[32] S. G. Shandilya and M. Timme, "Inferring network topology from complex dynamics," *New Journal of Physics*, vol. 13, no. 1, p. 013004, 2011.

[33] J. P. Bagrow, S. Desu, M. R. Frank, N. Manukyan, L. Mitchell, A. Reagan, E. E. Bloedorn, L. B. Booker, L. K. Branting, M. J. Smith *et al.*, "Shadow networks: Discovering hidden nodes with models of information flow," *arXiv preprint arXiv:1312.6122*, 2013.

[34] S. W. Linderman and R. P. Adams, "Discovering latent network structure in point process data," *arXiv preprint arXiv:1402.0914*, 2014.

[35] H. Andersson and T. Britton, *Stochastic Epidemic Models and Their Statistical Analysis*. Springer New York, 2000.

[36] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Large deviations analysis of adaptive distributed detection," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 6112–6116.

[37] A. H. Sayed, S.-Y. Tu, J. Chen, X. Zhao, and Z. J. Towfic, "Diffusion strategies for adaptation and learning over networks: an examination of distributed strategies and network behavior," *Signal Processing Magazine, IEEE*, vol. 30, no. 3, pp. 155–171, 2013.

[38] P. Braca, S. Marano, V. Matta, and A. H. Sayed, "Asymptotic performance of adaptive distributed detection over networks," *arXiv preprint arXiv:1401.5742*, 2014.

[39] F. W. Crawford, "The graphical structure of respondent-driven sampling," *ArXiv: 1406.0721*, 2015.

[40] O. Frank and D. Strauss, "Markov graphs," *Journal of the American Statistical Association*, vol. 81, no. 395, pp. 832–842, 1986.

[41] S. Wasserman and P. Pattison, "Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*," *Psychometrika*, vol. 61, no. 3, pp. 401–425, 1996.

[42] T. A. Snijders and M. A. Van Duijn, "Conditional maximum likelihood estimation under various specifications of exponential random graph models," in *Contributions to social network analysis, information theory, and other topics in statistics: A Festschrift in honour of Ove Frank*, J. Hagberg, Ed. Stockholm, Sweden: University of Stockholm, Department of Statistics, 2002, pp. 117–134.

[43] A. S. Klovdahl, J. J. Potterat, D. E. Woodhouse, J. B. Muth, S. Q. Muth, and W. W. Darrow, "Social networks and infectious disease: The Colorado Springs study," *Social Science & Medicine*, vol. 38, no. 1, pp. 79–88, 1994.

[44] D. E. Woodhouse, R. B. Rothenberg, J. J. Potterat, W. W. Darrow, S. Q. Muth, A. S. Klovdahl, H. P. Zimmerman, H. L. Rogers, T. S. Maldonado, J. B. Muth *et al.*, "Mapping a social network of heterosexuals at high risk for HIV infection," *Aids*, vol. 8, no. 9, pp. 1331–1336, 1994.

[45] R. B. Rothenberg, D. E. Woodhouse, J. J. Potterat, S. Q. Muth, W. W. Darrow, and A. S. Klovdahl, "Social networks in disease transmission: the Colorado Springs study," in *Social Networks, Drug Abuse, and HIV Transmission*, R. Needle, S. G. Genser, and R. T. Trotter, Eds. US Department of Health and Human Services, Public Health Service, National Institutes of Health, National Institute on Drug Abuse, 1995.

[46] V. E. Krebs, "Mapping networks of terrorist cells," *Connections*, vol. 24, no. 3, pp. 43–52, 2002.

[47] R. Lindelauf, P. Borm, and H. Hamers, "The influence of secrecy on the communication structure of covert networks," *Social Networks*, vol. 31, no. 2, pp. 126–137, 2009.

[48] R. Goyal, J. Blitzstein, and V. de Gruttola, "Sampling networks from their posterior predictive distribution," *Network Science*, vol. 2, no. 01, pp. 107–131, 2014.