ISSN 2278-3091
Volume 8, No.1.1, 2019
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse5081.12019.pdf
https://doi.org/10.30534/ijatcse/2019/5081.12019

Extracting Information Based on Partial or Complete Network Data

James Carbaugh[1], Matthew Fletcher[1], Ralucca Gera[1], Woei Chieh Lee[1], Russell Nelson[1], Joyati Debnath[2]
[1]Applied Mathematics Department, Naval Postgraduate School, Monterey, CA 93943
[2]Dept. of Mathematics and Statistics, Winona State University, Winona, MN 55987
jcarbaugh@nps.edu, mfletcher@nps.edu,rgera@nps.edu, wchieh@nps.edu, rnelson@nps.edu, jdebnath@winona.edu

ABSTRACT

General frameworks that works for all types of networks usually do not produce reliable results. The frame- work in question here is extracting information from large unknown networks encountered in the real world. Generally, researchers and operators work with incomplete data. Under- standing and particularly measuring a network's structure is a complex problem, and there is no general reliable way of measuring the structure in order to compare networks. In this research we present heuristic methods to gather information from a given network. We introduce an algorithm to place those monitors according to betweenness, closeness, degree and a new centrality called 2-hop centrality.

Key words: Information, Network, Random, Complex

1. INTRODUCTION

Complex networks are used to describe a myriad of interactions or affiliations such as organizational structures and relationships in social, informational, technological or biological systems. A network's topology helps in under- standing and studying complex networks and is challenging due to the large number of non-isomorphic networks for a given number of nodes. The purpose of this paper is to utilize an optimal number of monitors to infer a network. The research focuses on how to place the monitors in the network with the goal of observing as much of true network as possible. Specifically, we desire to add the minimum number of monitors to a network such that 90% of edges are inferred.

Inferring of complex network is done with the knowledge of the network at a random starting node, or with partial information collected from network devices (such as knowledge of some of the nodes present in the network), or with complete information (in which case one could use the current knowledge to further monitor the network, or to re-infer an evolving network).

Bliss, Danforth and Dodds [1] present techniques of infer- ring the topology of complex networks. These techniques are based on sampling nodes, sampling edges, the exploration of networks using random walks, or snowball sampling based on chain referral sampling [2], [3].

Although techniques such as *Random edge selection* and *depth and breadth first search network traversal* do not perform well overall; *simple uniform random node selection* performs very well. The best performing methods are the ones based on random-walks starting at an arbitrary seed node (with the added probability of p at each node to jumpout of the random walk to the seed node or another arbitrary node) [3].

The *k-Vertex Maximum Domination*, introduced by Miyano and Ono in [4], is the parameter that gives the ideal placement of monitors to infer all the nodes. k-Vertex Maximum Domination (k-MaxVD) finds a subset of the nodes with size k that maximizes the cardinality of dominated nodes (DN). That is, maximize $U_{v \in DN} N[V]$ Note that this optimization may produce a dominating set for some values of k, although generally not all the nodes in the network are dominated. The authors show that a simple greedy strategy achieves an approximation ratio of $1-\frac{1}{e}$ for k-MaxVD, and this approximation ratio is the best possible for k-MaxVD unless P = NP.

Haddadi et al. [5] published a survey on the research on Internet's network topology over the past decade that studied the techniques for inference, modeled and generated Internet topology at both the router and administrative levels. They also presented the mathematical models assigned to various topologies using traceroute.

A new research area, [6] namely network tomography assumes access to a few end-nodes in the network and the path of communication between them in order to study internal characteristics of the Internet. Malekzadeh and MacGregor [7] present an overview of these techniques, particularly showing the two popular approaches: *construc- tive algorithms* that create binary tree topology by starting with leaves, and *maximum-likelihood* approaches based on Bayesian estimator [8] and on Markov Chain Monte Carlo procedures [9].

Other current techniques not necessarily using complex networks are based on differential equations given one obser- vation of one collective dynamical trajectory [10], statistical dependence between observations [11], as well as machine learning based on frequency of small subnetworks [12].

2. DEFINITIONS

For the current research, a N_k -monitor is introduced, which when placed on a node it discovers: (a) all vertices within distance k of the monitor, (b) all edges between the monitor and the neighbors at each step, and (c) the edges between all i-step neighbors from the monitor ($1 \leq i \leq k-1$).

Definition 1. *Traditionally, a node monitors its neighbors only. We extend it to monitor all the nodes and edges that are at distance k or less from it: a node i* ***k-step-monitors*** *another node j if* $d(i, j) \leq k$. *Then a node i also* ***k-step-monitors*** *an edge* jL *if either* $d(i, j) < k$ *or* $d(i, L) < k$. *Such a node i is called an* N_k ***-monitor node.***

The 0-step-monitor just discovers itself. The 1-step-monitor discovers all of its neighbors and the edge to its neighbors, which matches the traditional definition of degree, which then can be used a guiding factor in placing monitors. Thus we study the case of $k = 2$, which defines the N_2-monitor, as depicted in Figure 1.

Figure. 1: A network and a monitor placed at node i

Similar parameters, such as 2-step domination and distance-2 domination below, have been used in graph theory. For a network tt, the 2-step neighborhood of a node $i \in V(G)$ is $N_2[i] = \{j : d(i, j) \leq 2\}$ (which includes i itself), $\forall i \in V(G)$. For an arbitrary node $i \in V(G)$ the 2-step neighborhood $N_2[i]$ is seen by an N_2-monitor placed at i. Also, a set $S \subseteq V(G)$ is a k-step dominating set if for every node $i \in V(G) - S$, there exists a path of length k from i to some node in S [13]. The k-step domination number, $\gamma_{\leq k}(G)$, is the minimum cardinality of a k-step dominating set of G, where a distance-k dominating set S is a subset of the nodes such that $\forall i \notin S, \exists j \in S$ such that $d(i, j) \leq k$. In the case of $k = 2$, both of these parameters are very close to the N_2-monitor number, but distinct since a node in a distance-2 dominating set doesn't necessarily dominate its neighbors.

For the purpose of this paper we use the following centralities based on the node's importance: betweenness centrality, closeness centrality, degree centrality, eigenvector centrality [14] and communicability. The *communicability centrality*, or subgraph centrality, of a node i is the sum of closed walks of all lengths starting and ending at node i. $SC(i) = \sum_{j=1}^{n} (v_j^{i-th})^2 e^{\lambda_j}$, where v_j^{i-th} is the i-th entry of the an eigenvector v_j corresponding to the eigenvalues λ_j [15]. The *graph density* is defined as $D = \frac{2|E|}{|V|(|V|-1)}$.

3. METHODOLOGY AND NETWORK DESCRIPTION

For monitoring/observing the network monitors are placed on the nodes with the highest centrality at first. First, best centrality-based method is to be determined for the placing of the minimum number of monitors in two scenarios. One, when all information of the network to be inferred is given. This provides insights into which centralities provide the best monitor placement methodology when given maximum information regarding the network in question. Two, when only partial information of the network to be inferred is given, yet the goal is still to monitor the whole network from this partial information.

A. Networks

We describe the networks analyzed, show an overview of each in Table 1 followed by a short description.

Table 1: The Order and Size of the Studied Networks

Network	Order	Size	Density
Collaboration	5242	14,496	0.00105
Erdós - Rényi	5242	14,496	0.00105
Barabási - Albert	5242	15,717	0.00114
Facebook	4039	88,234	0.0108
Configuration	4039	85,643	0.0105

Collaboration Network: The General Relativity Collaboration Network from SNAP [16], captures the collaboration between authors who submit papers to the General Relativity and Quantum Cosmology category, whose undirected edges capture co-authorship of papers.

Erdós Rényi Network: The Erdós Rényi Random Network was created using the Python's random network model $ER(n, m)$ to be comparable to the Collaboration Network.

Barabási-Albert Network: The Barabási Albert Network was created using the Python's preferential attachment model $BA(n, d)$ also to be comparable to the Collaboration Network.

Facebook Network: The undirected Facebook Network from SNAP [16] includes people, circles and ego networks. The information on the people has profile information, while the circles are the friend lists. The ego network consists of the people' friend connections [17].

Configuration Model Network: The Configuration Model Network was created using the Python's Configuration model and the degree sequence of the Facebook Network.

B. Perfect Information Derivation

The investigation began with the situation in which complete information about the network is known. Then monitors are placed on the nodes with the highest centrality value (for each centrality). Our analysis shows the marginal benefit of each percent of monitors added, as the number of monitors increased from no monitors (0%) to half of the network's nodes with monitors placed on them (50%). We then explore the usefulness of different centralities to see at what point we get 90% of the edges. The algorithm used for perfect information for the data sets is presented in Algorithm 1, and the results and their analysis in Section 4.

C. Partial Information Derivation

A more realistic scenario, is the case where only partial information about the network is available, as rarely we

Algorithm 1 Algorithm for perfect information.

for each centrality type **do**
  Calculate centrality for all nodes
  Sort by centrality values in nonincreasing order
  for each num of Monitor **do**
    Place given num of Monitors on nodes with highest centrality value
    List edges inferred by the N_2 monitors as given in Definition 1
    Store edges inferred into results
Save results to file

know the whole network as it evolves or we are not privileged to that information. To simulate partial information scenario, we randomly sample a set of nodes from the network, and generate a subnetwork with the edges induced by the selected nodes. This induced subnetwork is partial information network, as we formally describ it next.

Given a network, $G = (V, E)$, let $p \in [0, 1]$ be the information coefficient, where $p = 0$ denotes no information on G and $p = 1$ denotes perfect information on G. We create an partial information network, G^r, by $|V^r| = \lceil p \cdot V \rceil$, and $G^r = G[V^r]$, where the partial information network, G^r is the subnetwork induced by the selected nodes $V^r \subseteq V(G)$.

Algorithm 2 Algorithm for partial information.

for trials = 1 to 10 **do**
  Randomly pick 50% of nodes and create the induced subgraph formed by selected nodes
  for each centrality type **do**
    Calculate centrality for all nodes in random induced subgraph
    Sort by centrality values in nonincreasing order
    for each num of Monitor **do**
      Place given num of Monitors on nodes with highest centrality value on true network
      List edges inferred by the N_2 monitors as given in Definition 1
      Store edges inferred into results
Save results to file

In this paper $p = 0.5$ is chosen in order to generate a partial information network G^r with half of the nodes of the original network. Then a subnetwork is induced by adding all the existing edges between these nodes, and centralities are evaluated. Then monitors are placed on nodes in order starting with the highest centrality values. After each monitor placement, the evaluation of the monitored network to the original network is plotted. Of course, the value of p can be varied as desired. When p is lower, less information is available.

Due to the random selection of nodes to produce a subnetwork the process is replicated ten times for each selection of network and each centrality. Then the mean of the ten results for each selection of network and centrality are analyzed. The algorithm for the process is shown in Algorithm 2 and the results and their analysis in Section IV.

4. RESULTS AND ANALYSIS

For each of the partial and perfect information scenario, a table that shows the percent of nodes monitored for finding the minimum number of monitors (that could detect 90% of a network's edges) is presented. Then, the percent of edges inferred as a function of the percent of nodes monitored are shown graphically. The plots and their interpretation provide relative performance of each centrality across a wide range of percent monitor placement.

A. Perfect Information Results

We first consider the scenario when the whole network is given to be monitored. Table II presents the overview of the percent of nodes that have monitors on them in order to discover 90% of the edges of each network.

Table 2: Percent of Node Monitor Placements Needed to Infer 90% of Edges Using Perfect Information

Centrality	Collaborati	E-R	B-A	FB	Confi
Betweenness	10%	11%	2%	1%	1%
Closeness	50%	14%	3%	1%	1%
Degree	19%	11%	2%	1%	1%
Eigenvector	44%	11%	3%	30	1%
Communicabi	30%	12%	3%	29	1%

In each of our centrality-based methods, monitors are placed according to rank order. At the 90% benchmark (depicted as a dashed line in the plots) as represented in Table 2, it appears that betweenness dominates all other centralities. Degree centrality is a very close second, and is only worse than the betweenness centrality in the collaboration network. The communicability and eigenvector are third and fourth best, and closeness is consistently the worst performer. The configuration model network provides no information to distinguish metrics, while the collaboration network has the greatest variance among its results.

B. Partial Information Results

In this subsection, we consider the scenario when part of the network is allowed to be observed, before monitoring the whole network. Table III presents general information regarding the data sets which are described in more details. Decisions are made on where to place the monitors based on knowing just half the nodes in the network (and all the edges between these nodes). The performance of the centrality-based method changes less across networks than the cost of monitor placement by network type. With partial information, degree or betweenness is the best centrality-based method, depending on the network. communicability and eigenvector again come in third place, with closeness consistently performing worst. The relative ranking of centrality methods is almost exactly the same, whether initially the entire network or half the network is accessed.

Table 3: Percent of Node Monitor Placements Needed to Infer 90% of Edges Using Partial Information

Centrality	Collabor	E-	B	F	Conf
Betweennes	18%	17	5	2	1%
Closeness	31%	19	9	33	1%
Degree	20%	16	5	7	1%
Eigenvector	30%	19	9	30	1%
Communica	30%	17	7	33	1%

In the Collaboration network, closeness and eigenvector perform better with partial information being exposed. This is the trend for both centralities at every step of the inference and is due to the spread of the nodes throughout the network, that prevents placing monitors in the same components based on the centrality in each component.

C. Marginal benefit analysis

At a more detailed level, the marginal benefit analysis of placing each percent of monitors is analyzed. Comparison is made based on percent of nodes occupied by monitors since the networks are not of the same size.

The overview of the partial information with the perfect information is compared in order to see if any centrality-based methods perform consistently regardless of the amount of information with which the network is approached. Additionally, the cost of monitor placement (that is, the number of monitors we must place to achieve the desired inference) changes by the nature of the network, regardless of centrality method.

The results of the collaboration network are shown in Figure 2 and Figure 3. The performance of each centrality varies greatly at the 90% benchmark. By design, the Collobotation network and the Erdos Renyi graph have the same density, the performance of each centrality method on the network varies widely. Density itself does not determine the performance of the centrality based metrics. The collaboration network is unique in the study because the collaboration network is a disconnected graph, containing over 300 components. The disconnected nature of the graph underscores a fundamental challenge for any centrality-based inference metric. That is, since all of the centralities are based on the edges of the network, if there is no edge between two nodes, no edge-based metric will ever detect the next component from a previous one.

The one exception to the reduced efficiency of monitor placement is the case of the closeness-based method in the collaboration network. Recall that monitors are placed according to their rank in a sorted list of nodes with highest closeness down to lowest closeness. With perfect information, this list of nodes is fixed. In this simulation, in each trial the list of nodes in rank of closeness is unique. It is possible, then, that as nodes from this list are dropped or rearranged, monitors are placed at nodes which are more spread out throughout the network. In this case, monitors overlap less and become more efficient in inferring the network.

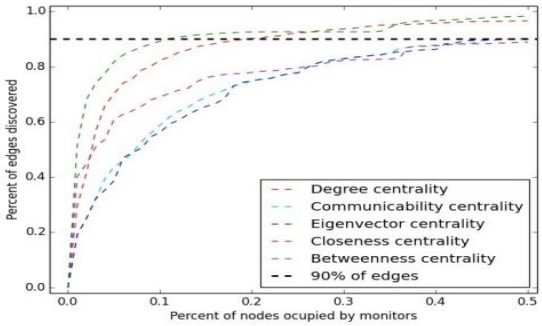

Figure. 2: Complete information for Collaboration Networks

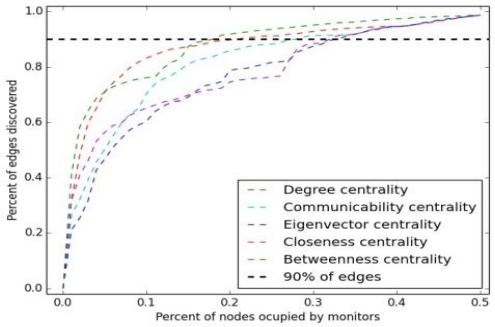

Figure. 3: Partial information for Collaboration Networks

The results for Erdos Renyi are in figure 4 and figure 5. In the three synthetic networks, less variation between the different choices of centralities is seen, as compared to the real data.

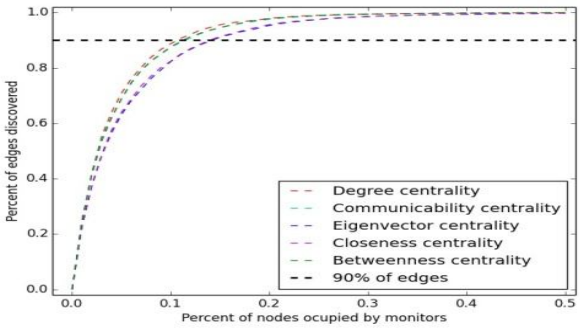

Figure 4: Complete information for Erdos Renyi

Recall that Barabasi- Albert network and Erdos Renyi network are of the same order. In other words, the placement of k monitors will achieve the desired result faster on a network of preferential attachment than a random graph because the k best monitors are more important in Barabasi-Albert network that k -best in the Erdos-Renyi network.

The Barabasi- Albert network and Erdos Renyi network follow a similar pattern in terms of centrality perfor- mance (see Figures 4 and 6). The ability of the nodes of

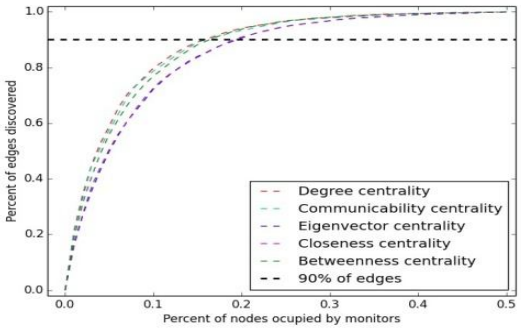

Figure 5: Partial information for Erdos Renyi

highest centrality to serve as monitors is much greater in a network of preferential attachment than in a random graph. This result comes from the key difference between the two graphs. That is, in a network of preferential attachment, certain nodes are extremely important by design. These nodes are the nodes that have "grown" first.

Recall that the density of the random graph and the density of the collaboration network are the same. Additionally, the density of the random network and Barbasi Albert network do not differ greatly, yet monitor placement results differ widely. It appears that graph density itself does not necessarily determine the performance of centrality-based monitor placement. The degree distribution of the network itself may give greater insight into the performance of N_2 monitors than graph density.

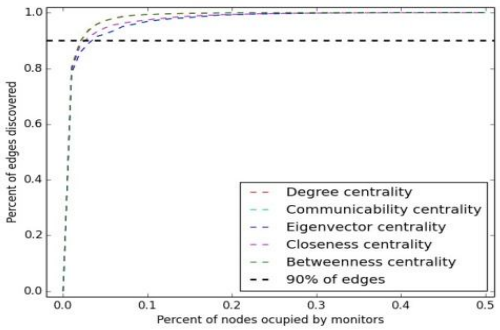

Figure 6: Complete information Baranasi-Albert

In the Facebook network's data of Figure 8 and 9, there are several nodes of extremely high degree. When started with partial information, the algorithm does not show some of these high degree nodes when other centralities are considered. The effect of losing some of these very important nodes is felt across two of the centrality methods.

Notice in Figures 8 and Figure 9 that the Facebook network monitor placement can be anywhere from least expensive to most expensive depending on the centrality-based method. That is, it can be inferred with 1% of the nodes using all but the communicability and eigenvector

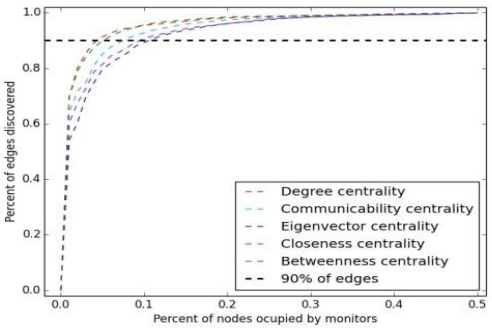

Figure 7: Partial information for Barabasi-Albert

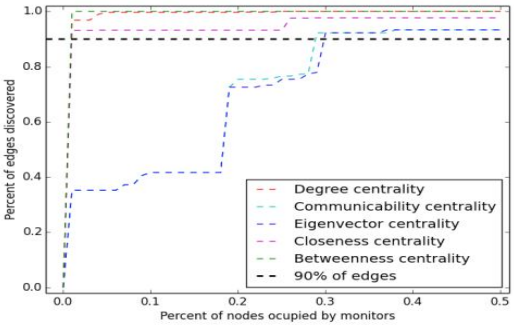

Figure. 8: Complete information for Facebook Networks

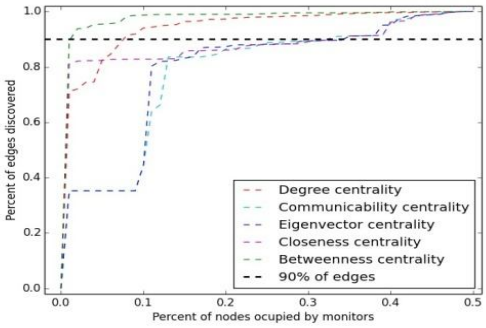

Figure. 9: Partial information for Facebook Networks

centrality, yet it takes almost 30% of the nodes if these high degree nodes are missed. When monitors were placed on this network using communicability, the first monitor discovered 35.2% of the network, and by the time 9% of nodes have monitors, the percent of edges discovered has not changed. The network is being discovered slowly starting at 14% of nodes with monitors, showing that important nodes according to these two centralities share most of their neighbors and edges leading to the neighbors.

Since only 1% of nodes are required to infer 90% of the edges for all different centrality based algorithms, the configuration network provides no information with which to differentiate the relative performance of the centralities,

nor between complete versus partial information, as shown in Figure 10. This is because the configuration network is denser that the Erdos-Renyi network and it has the hubs that make the inference process extremely efficient.

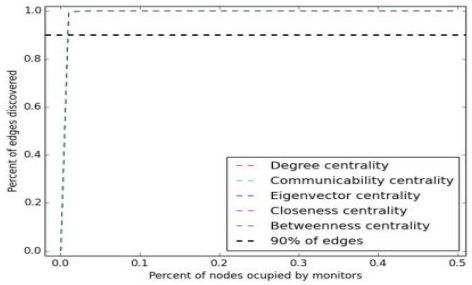

Figure . 10: Complete information for Configuration Model

5. CONCLUSION

In this paper a framework for network inference is created based on the concept of centrality. Our goal is defined as monitoring 90% of the edges of a true network in an inferred network, and a marginal benefit per percent monitor utilized is presented. Inference solely in terms of edges without regard to nodes inferred is described. Given a particular network, each node of the network is ranked by its centrality and placed monitors sequentially according to each node's rank. The percent of the network inferred is calculated as a function of monitor placement for two examples of classical graphs and two real world networks. Since there is no need for inference of a network when perfect information is available, the process is expanded to a more realistic scenario when started with limited information.

Notice that placing monitors according to degree or betweenness centrality is consistently the best centrality-based method for monitor placement, even in the presence of partial information. Creating the subnetwork of partial information is computationally expensive, and sorting these subnetworks' nodes according to their centrality is even more computationally expensive. Therefore, using the degree centrality-based method is recommended because its performance is computationally inexpensive and robust to limited information across network types.

With perfect information (that is complete knowledge of the network) 90% of the networks edges can be monitored by using between 1% and over half of the nodes on the studied networks, and an analysis of this wide range is presented. In addition to the complete knowledge of the networks, the algorithms are tested against a subnetwork of the true networks that consists of 50% of the original networks nodes and the edges between these nodes. It is shown that even without perfect information, a network's edges using between 1% and 33% of the nodes of the true network as monitors can be effectively detected. An analysis of which centralities produce a better ranking of the nodes for network inference is also presented. This study serves to validate the process of placing monitors according to degree centrality in more sophisticated algorithms developing or in development. This method of degree centrality-based monitor placement is efficient and robust to network type and limited information.

ACKNOWLEDGMENT

Part of the current research was sponsored by DoD.

REFERENCES

[1] 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. https://doi.org/10.1371/journal.pone.0108471

[2] P. Biernacki and D. Waldorf, "Snowball sampling: Problems and tech- niques of chain referral sampling," Sociological methods & research, vol. 10, no. 2, pp. 141–163, 1981. https://doi.org/10.1177/004912418101000205

[3] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2006, pp. 631–636. https://doi.org/10.1145/1150402.1150479

[4] E. Miyano and H. Ono, "Maximum domination problem," in Proceedings of the Seventeenth Computing: The Australasian Theory Symposium-Volume 119. Australian Computer Society, Inc., 2011, pp. 55–62.

[5] H. Haddadi, M. Rio, G. Iannaccone, A. Moore, and R. Mortier, "Network topologies: inference, modeling, and generation," Communications Surveys & Tutorials, IEEE, vol. 10, no. 2, pp. 48–69, 2008. https://doi.org/10.1109/COMST.2008.4564479

[6] Y. Vardi, "Network tomography: Estimating source-destination traf- fic intensities from link data," Journal of the American Statistical Association, vol. 91, no. 433, pp. 365–377, 1996. https://doi.org/10.1080/01621459.1996.10476697

[7] A. Malekzadeh and M. H. MacGregor, "Network topology inference from end-to-end unicast measurements," in Advanced Information Networking and Applications Workshops (WAINA), 2013 27th International Conference on. IEEE, 2013, pp. 1101–1106. https://doi.org/10.1109/WAINA.2013.215

[8] N. G. Duffield, J. Horowitz, F. Lo Presti, and D. Towsley, "Multicast topology inference from measured end-to-end loss," Information Theory, IEEE Transactions on, vol. 48, no. 1, pp. 26–45, 2002. https://doi.org/10.1109/18.971737

[9] F. W. Crawford, D. Zelterman, W. H. Mulder, M. A. Suchard, P. M. Aronow, A. Coppock, D. P. Green, R. E. Weiss, and V. N. Minin, "The graphical structure of respondent-driven sampling," arXiv preprint arXiv:1406.0721, 2014.

[10] S. G. Shandilya and M. Timme, "Inferring network topology from complex dynamics," New Journal of Physics, vol. 13, no. 1, p. 013004, 2011. https://doi.org/10.1088/1367-2630/13/1/013004

[11] K. Tieu, G. Dalley, and W. E. L. Grimson, "Inference of non- overlapping camera network topology by measuring statistical de- pendence," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2. IEEE, 2005, pp. 1842–1849.

[12] M. Middendorf, E. Ziv, and C. H. Wiggins, "Inferring network mech- anisms: the drosophila melanogaster protein interaction network," Proceedings of the National Academy of Sciences of the United States of America, vol. 102, no. 9, pp. 3192–3197, 2005. https://doi.org/10.1073/pnas.0409515102

[13] J. Boland, T. Haynes, and L. Lawson, "Domination from a distance," Congr. Numer., vol. 103, pp. 89–96, 1994.

[14] M. Newman, Networks: An Introduction. New York, NY, USA: Oxford University Press, Inc., 2010. https://doi.org/10.1093/acprof:oso/9780199206650.001.0001

[15] E. Estrada and J. A. Rodriguez-Velazquez, "Subgraph centrality in complex networks," Physical Review E, vol. 71, no. 5, p. 056103, 2005. https://doi.org/10.1103/PhysRevE.71.056103

[16] J. Leskovec and C. F. Andrej Krevl, "Graph evolution: Densification and shrinking diameters. acm transactions on knowledge discovery from data (acm tkdd), 1(1)," http://snap.stanford.edu/data, Jun. 2007.

[17] J. McAuley and J. Leskovec, "Discovering social circles in ego networks," ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 8, no. 1, p. 4, 2014. https://doi.org/10.1145/2556612