A Robust Quality Metric for Noisy Incomplete Networks

Suhansanu Kumar*
Dept. of CS,
UIUC, IL

Soumya Sarkar*
Dept. of CSE,
IIT Kharagpur, India

Sanjukta Bhowmick
Dept. of CS,
University of Nebraska, Omaha

Animesh Mukherjee
Dept. of CSE,
IIT Kharagpur, India

*Both the authors have equal contributions.

***Abstract*—Network analysis has become an ubiquitous tool for studying large-scale datasets. However, networks often do not provide an exact representation of the underlying data. They can be noisy, that is, have extra or missing edges compared to the actual relations in the data. Additionally, smaller sampled subnetworks of larger networks are used in many applications, so that the analysis can be performed in a reasonable time. Measuring the robustness of noisy networks and ensuring the accuracy of results in incomplete networks is an important problem in network analysis. Although both these problems are associated with accuracy, to date they have been considered as two separate aspects.**

In this paper, we present permanence as the first unifying metric to handle different accuracy aspects of noisy, incomplete networks. Specifically, we show that (i) permanence is an excellent indicator of how a network changes under noise, (ii) vertices with high permanence are very effective seeds from message broadcast and (iii) sampling based on high permanence vertices can retain the cluster structure of the network. Our experiments on comparing permanence with other standard metrics and algorithms over a set of real-world and synthetic networks demonstrate the advantage of using permanence as a robust metric for evaluating and analyzing noisy incomplete networks.

I. Introduction

Network analysis has become an ubiquitous tool in understanding the properties of systems of interacting entities, such as those arising in bioinformatics [1], [2], social sciences [3], [4] and epidemiology [5]–[7]. The vertices in the network represent the entities in the system and the edges represent their pairwise interactions.

However, in the practical context, the network models are rarely an exact representation of the underlying system. This is because, most real-world networks are inherently noisy, i.e. they have missing or extra edges/vertices, as compared to the original interactions. Moreover, many networks are extremely large, and therefore, only smaller samples of the networks are analyzed to provide the answers within a reasonable time [1]. Thus, noise is due to limitations in data gathering or an incomplete view of the data (e.g., certain vertices and edges might remain unobserved due to privacy concerns); on the other hand, sampling is due to limitation in computational resources.

[1]Parallel algorithms can also be employed for large networks, but not all analysis algorithms have efficient parallel implementations. Furthermore, due to simultaneous operations, parallel algorithms can produce different results than sequential ones.

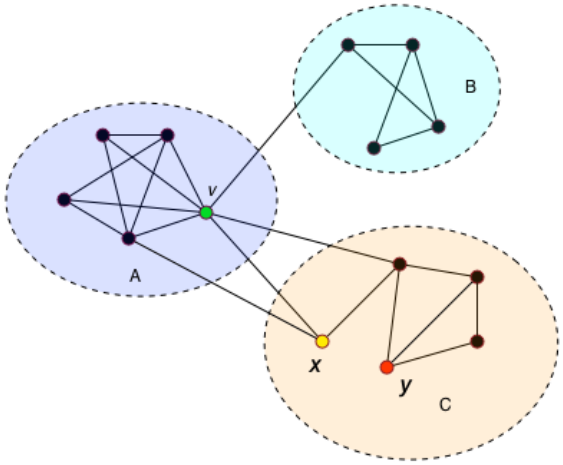

Fig. 1. Toy example depicting permanence of a vertex v (here, $I(v) = 4, D(v) = 7, E_{max}(v) = 2, c_{in}(v) = \frac{5}{6} \Rightarrow P(v) = 0.12$). While x and y both have same clustering coefficient $\frac{2}{3}$, they have different permanence values $P(x) = -\frac{5}{6}$ and $P(y) = \frac{2}{3}$.

The goal in noisy networks is to identify a network parameter, that can estimate the effect of noise, and indicate when the level of noise is too large to provide accurate analysis. The goal in network sampling is to find a metric based on which the network is sampled, so that important properties are preserved. Another problem related to incomplete networks is the spreading of information. Here the goal is to identify seed nodes, that can quickly propagate messages even when some edges are missing in the networks. To date, these three issues of evaluating noisy networks, spreading messages and sampling have been studied as different problems (see Section III for related algorithms).

In this paper, we provide a unified solution to these three problems. We posit that the net effect of all these problems are the same – we have to analyze a network where not all the edges are provided, and yet, guarantee reasonably accurate results. We need a metric that is sensitive to changes in the network, has a strong correlation with different network features, and, in turn, ensures higher robustness under sampling. Moreover, if this metric is reflective of large clusters in the

network, it would also enable effective seed selection for message spreading.

We present, *permanence*, as such a robust and accurate metric. Permanence was initially introduced as a metric for evaluating communities in [8]. However, because of the vertex-centric property of permanence as well as its relation to the modular structure of the network, it closely correlates with many intrinsic features of the network. The main contribution of our paper therefore is *to show that permanence serves as an ideal metric that can be effectively used for measuring and ensuring robustness in noisy incomplete networks.* While the different components of permanence are appropriately sensitive to noise, the ranking of the nodes based on permanence remain relatively more stable over varying levels of noise thus making it more robust than the state-of-the-art metrics.

We support our claim through a wide range of experiments that compare permanence with other metrics and algorithms. We show that compared to other standard network metrics, permanence is (i) more appropriately sensitive to noise (Section IV) and (ii) more closely correlated with different network properties (Section IV). We also show that (iii) high permanence vertices form effective seeds for fast broadcast of messages (Section 5), and (iv) sampling based on finding high permanence of vertices and their subgraphs, preserves network properties better compared to other sampling algorithms (Section 6). The last two advantages hold true even when the sampling is performed on noisy networks.

The rest of the paper is organised as follows. In the next section we define permanence as a quality metric for noisy incomplete networks. In section III, we outline the datasets, the noise models as well as the baseline methods for spreading and sampling. In section IV, we outline experiments to show that permanence is appropriately sensitive to noise compared to other test metrics. In addition, we show that permanence is highly correlated to network structural properties. In the next section, we present results from the spreading experiments and compare the performance of permanence with that of the other baselines. In section VI we outline the results of the sampling experiments. In section VII, we discuss the key empirical properties of permanence that makes it an ideal quality metric for noisy networks. In section VIII we present a comprehensive review of the state-of-the-art approaches. Finally, we conclude and summarise future directions in section IX.

II. Permanence as a centrality metric

We first provide a brief introduction to permanence. Permanence was introduced in [8] as a vertex-centric metric for evaluating the quality of communities in a network. If the community structure is known, permanence is defined for each vertex as follows; $P(v) = [\frac{I(v)}{E_{max}(v)} \times \frac{1}{D(v)} - (1 - c_{in}(v))]$ where $I(v)$ is the number of neighbours of v in its community, $E_{max}(v)$ is the maximum number of connections of v in its neighbouring community, $D(v)$ is the degree of v and $c_{in}(v)$ is the internal clustering coefficient of the internal nodes of v in the community. Permanence tries to formulate the "pull" experienced by a vertex v from its external neighbors, i.e., neighbors residing in communities different from that of v. If this pull is low (low $E_{max}(v)$) and the internal neighbors in the community of v are tightly connected (high $c_{in}(v)$) then there is a high probability that v is very "'stable" in its own community. In other words, the community assignment for v is correct.

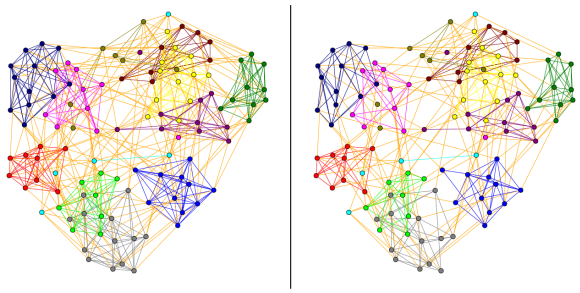

Fig. 2. Football network with ground truth community shown in palette of rainbow colors. On the left is the original network while on the right is the same network with 30% of the edges removed uniformly at random.

The value of permanence ranges from 1 (all neighbors form a cluster and are in the same community as the vertex) to -1 (all neighbors form a cluster and are in a different community than the vertex). Vertices in singleton communities have permanence zero. The permanence of a network is the average permanence of all its vertices. Therefore a network with more community-like structure will have higher permanence. For instance, the permanence of the node v in Fig. 1 is equal to 0.12.

Since it is vertex-centric, permanence can also be considered as a centrality metric, where vertices in tightly bound clusters are marked to be important. Nodes with high permanence indicate that their neighbors form cliques or near-cliques. While clustering coefficient can also provide this information, it considers the entire set of neighbors around a node v. In contrast, permanence uses a restricted set, consisting only those neighbors that are part of the same community as v. This difference between ordinary clustering coefficient and permanence is illustrated in Fig. 1; consider the two nodes x and y both of which have exactly same clustering co-efficient but different values of permanence. Therefore a high clustering coefficient does not indicate whether a vertex strongly belongs to a community, whereas a high permanence means that the vertex belongs strongly within a community.

Permanence therefore provides a unique centrality value, that not only indicates the ranking of a vertex as per its importance, but also provides information about its position in a community. Because they belong to the core of communities, high permanence vertices are very effective as seeds for broadcasting messages and also as good bases for sampling. Moreover, because these vertices belong to tightly connected clusters their overall rankings are less affected by the varying levels of noise in the system. Therefore message spreading and sampling using high permanence vertices is more resistant to noise within the network

Interestingly, the mean permanence over *all* the vertices is very sensitive to the effect of noise. In Fig. 2, we illustrate

this phenomena for the structure of the well-known football network [9] (see Section III for description). The left network shows the community structures present in the original network (each community is represented by a different color), while the right network shows the structure when 30% of the edges have been removed uniformly at random. Note that the community labels for such moderate levels of perturbation to the network do not almost change. Further, while the standard quality metrics like modularity, conductance and cut-ratio cannot signal the differences between the left and the right network, permanence is the only metric that signals this difference. While from left to right network, average permanence significantly declines from 0.266 to 0.071, the other metrics hardly change – modularity changes from 0.553 to 0.557, conductance changes from 0.402 to 0.399 and cut-ratio changes from 0.039 to 0.028[2]. Thus mean permanence over the network is very appropriately sensitive to noise as compared to the other metrics.

III. Experimental setup

In this section, we provide an overview of the datasets used, the different noise models and levels, comparative evaluation metrics and competing algorithms used for spreading and sampling processes.

A. Dataset

In our experiments, we consider a multitude of networks – both synthetic and real world for evaluation of different metrics wherever appropriate.

For analysing the effects of different forms of noise as well as for the spreading related experiments we use the following networks:

LFR Benchmark: We use the benchmark **LFR model** [10] that generates different networks and ground-truth communities of various quality. The system allows one to set the number of nodes (n), the average ($\langle k \rangle$) and the maximum degrees (k_{max}), the degree distribution and the community size distribution exponents. For our purposes, we use $n = 1000$, keeping all other parameters to their default values as specified by the authors in their implementation[3]. The only parameter that we vary is the mixing co-efficient (μ) which represents the goodness of the communities (average ratio of interal edges of a node to its degree). The values that we choose for μ are 0.1 and 0.3 as appropriate. Note that the lower the value of μ the better are the communities.

Football: This is a small real world dataset from Girvan et al. [9] which represents a **football network** based on the American football games between division IA colleges during regular season fall of 2000. The vertices in the graph represent teams (identified by their college names), and edges represent regular-season games between the two teams they connect. The teams are known to belong to "conferences" containing around 8 to 12 teams each and there are more games played within a conference than in between conferences. Each conference is therefore assumed to be a ground-truth community in this network.

For sampling, we conduct our experiments on large real world networks where the primary objective is to obtain a representative sample as close as possible to the original network (even when only a noisy version of the original network is available).

ca-CondMat and ca-HepPh: High energy physics, phenomenology (ca-HepPh) and Condensed matter physics (ca-CondMat) are datasets taken from Leskovec et al. [11]. Both these datasets are collaboration networks from the e-print arXiv and covers scientific collaborations between authors of papers submitted in the respective categories. If an author i co-authored a paper with an author j, the graph contains a undirected edge from i to j. If the paper is co-authored by k authors this generates a completely connected (sub)graph on the k nodes. In this network, the ground-truth communities are not available; for computation of permanence we detect communities in the original network using the Informap algorithm [12], and use these as a proxy for the ground-truth. For the noisy versions of this network we assume that the obtained communities almost do not get altered (for the low to moderate noise levels that we investigate).

Email Enron: This network represents the email communication (edges) between different email addresses (nodes) which were originally posted by Federal Energy Regulatory Commission and made publicly available in [13]. We assume an un-directed version of the network and obtain the communities from the Infomap algorithm [12] to serve as a proxy for the ground-truth. Further, once again for the noisy versions of this network, we presume that the obtained communities remain almost unaffected.

We note the basic statistical properties of the above datasets in Table I.

B. Noise models

In this section, we describe the representative noise models, each of which attempts to emulate mechanisms by which noise could get introduced into the original network. These models are inspired by [14], developed primarily to simulate various real world sources of noise:

1) **Uniform noise model** – Often in real world networks, random edges might get missing owing to uniform uncertainty in the collection of data samples. A stochastic model that captures this kind of uniform uncertainty of the presence/absence of an edges defines the uniform noise model. To simulate this model we remove edges uniformly at random till a desired number of edges are removed. The number of edges removed is regulated by a pre-defined parameter (henceforth referred to as the noise level).
2) **Crawled noise model** — Crawling is one of the most important method of collection of online and offline network datasets – examples include crawling of online social networks [15], snowball sampling in sociological

[2]Note the scale-size is 2, 2, 1, 1 for permanence, modularity, conductance and cut-ratio respectively.

[3]https://sites.google.com/site/santofortunato/inthepress2

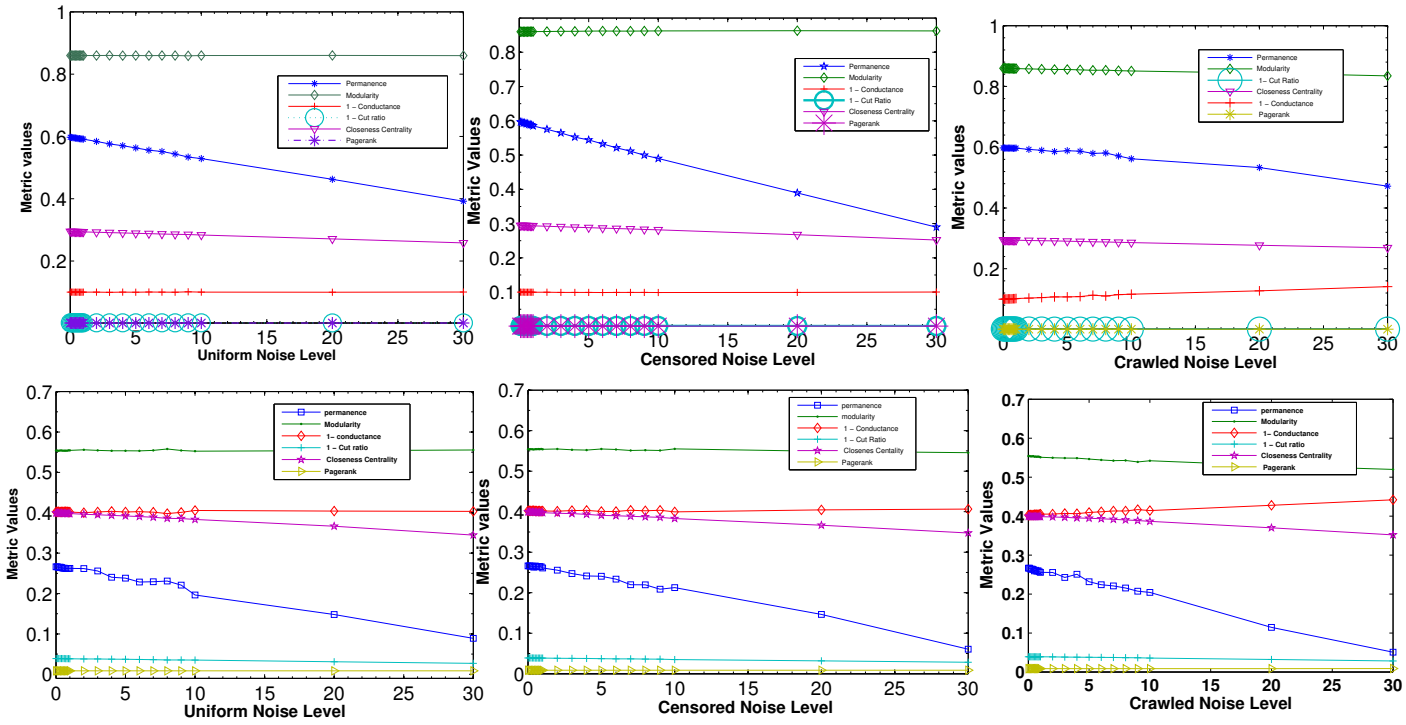

Fig. 3. Sensitivity of the different quality metrics for varying levels of noise. The top panel is for the LFR network ($\mu = 0.1$) and the bottom panel is for the football network.

Network	#Nodes	#Edges	<k>	k_{max}	$\|c\|$	n_c^{max}	n_c^{min}
Football	115	613	10.66	12	12	13	5
ca-CondMat	23133	93497	8.08	1383	2591	1226	1
ca-HepPh	12008	118521	19.74	281	1923	130	2
Email Enron	36692	183831	10.02	65	1101	162	1

TABLE I

DATASET STATISTICS. $|c|$ DENOTES THE NUMBER OF COMMUNITIES IN THE GROUND-TRUTH (EITHER AVAILABLE OR OBTAINED USING INFOMAP ALGORITHM [12]), n_c^{min} AND n_c^{max} DENOTE THE NUMBER OF NODES IN THE SMALLEST AND THE LARGEST SIZE COMMUNITIES RESPECTIVELY.

studies [16], boundary specification problems [17] and many others. In order to simulate this noise model, we start a BFS (breadth-first search) from the node that has the highest closeness centrality and collect edges till a pre-defined number of boundary edges are missed (determined by the noise level).

3) **Censored noise model** — Often while conducting surveys, there is a limit to the number of relationships that a respondent can answer (fixed choice problem). This puts a censorship on the degree of the node [18]. At each step, we randomly delete from the original network an edge of the highest degree node till a required number of edges (determined by the noise level) have been removed.

In specific, we are interested in how different baseline metrics get affected as compared to permanence due to the introduction of the above forms of noise in the original network. Some of these metrics are purely community-centric (since permanence is originally a community-centric design) such as modularity, cut-ratio, and conductance while others are some forms of centrality such as PageRank and closeness. Further, to keep our analysis as unbiased as possible, we do not allow formation of disconnected components while introducing noise.

C. Baseline spreading processes

In message spreading [19], a set of source vertices (initiators) start sending a message. At every time step, a vertex containing the message transfers the message to one of its neighbors who does not have the message. The algorithm terminates when all vertices have received the message. The selection of the initiators is critical to how quickly the message spreads. We identify the following baseline methods for selecting the initiator and compare the time required to spread the message in the noisy network (for various noise levels) with that of the case where initiators are selected based on permanence. Note the we assume the network community structure of the original network is known and it largely persists for the noisy networks also (since the noise levels we analyse is not extreme enough to completely destroy the communities).

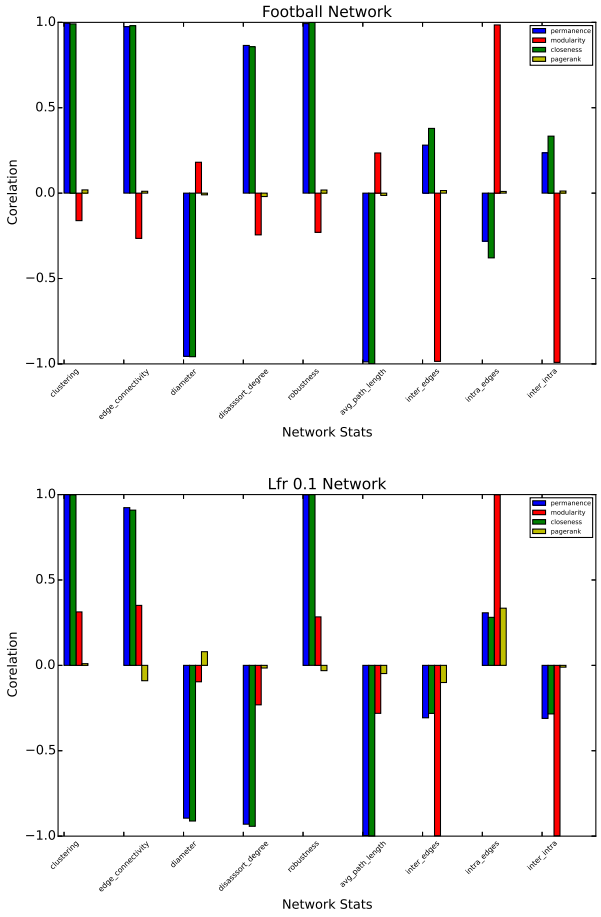

Fig. 4. Correlation of permanence, modularity, closeness centrality and PageRank with the different network features.

1) **Random selection:** From each identified community in the network a small fraction of initiators are randomly selected to pass the message.
2) **Degree based selection:** From each identified community in the network a small fraction of initiators are selected based on highest degree to pass the message.
3) **PageRank based selection:** In this case, to pass the message, we sample a small fraction of initiators from each identified community based on the highest PageRank.
4) **Closeness based selection:** In order to pass the message here, we sample a small fraction of initiators from each identified community based on highest closeness centrality.

D. Baseline sampling algorithms

In sampling task we consider a static graph and obtain samples of of varying sizes from the original graph. We also obtain samples from the noisy versions of the graph where the noise could be introduced through any of the noise models discussed above. The crucial difference among all existing sampling strategies is in the choice of the selection of a neighbor from the initial seed nodes which are selected randomly. We now discuss briefly the baselines which we use for sampling subnetworks.

1) **Random walk with restart (RWR):** We start at a random node i. The probability that we visit any neighbor w of i is proportional to the degree of that neighbor. At each step we have some probability to return to our starting node i (we consider this as 0.15 based on the literature).
2) **Metropolis Hasting Random Walk (MHRW):** RWR is biased towards high degree nodes. To eliminate this bias the idea of MHRW was introduced in [20] where at every iteration, from the current node v we randomly select a neighbor w and move there with probability $min(1, \frac{k_v}{k_w})$ where k_v and k_w are the degrees of the nodes v and w respectively. It can be shown that the resulting sample of nodes generated simulates a uniform distribution.
3) **Closeness:** Nodes are sampled by having a random walker traverse on the network preferentially based on closeness centrality values.
4) **PageRank:** Nodes are sampled by having a random walker traverse on the network preferentially based on PageRank values.

IV. Sensitivity of metrics under noise

Our first set of experiments focus on measuring how different network parameters behave as noise levels in networks are increased. Ideally, a *sensitive* parameter is one whose change is commensurate with the amount of noise applied. For small amounts of noise, the change in the parameter values should be low, whereas as the noise increases, the change should be much higher. *Our goal is to identify sensitive parameters, by computing which we can quickly determine whether a network significantly changed from its original topology.*

Methodology: We apply the three noise models on the synthetic and real-world benchmarks – LFR network ($\mu = 0.1$) and football network – from our test suite. For each increasing level of noise we compute the value of the candidate parameters. In addition to permanence, we select three community evaluating parameters, namely, modularity, cut-ratio and conductance. We also select two centrality metrics closeness centrality and PageRank. We compute the community-centric parameters based on the community assignment from the original network. Our rationale is that community detection is by itself expensive, so re-computing the community after each noise addition would defeat the purpose of quickly ascertaining the change in the network. Further, we have already motivated this issue with an example in Fig. 2 where we have seen that the community structure does not change much for tolerant levels of noise.

Results: Our experiments in Fig. 3 show that compared to all the other metrics permanence is the most sensitive to noise. Note that although the relative order of the values changes across different networks (for example, permanence has the second highest value in LFR ($\mu = 0.1$) and the fourth highest value in the football network), the effect of noise is similar

regardless of the network type or noise model. Almost all the other parameters remain constant, only permanence shows a clear decrease in values with increasing levels of noise.

A. Correlation of the parameters with network structure

The sensitivity of permanence indicates that this metric is more representative of network features. To test this hypothesis, we compute the correlation of these parameters with several different network structure features. The correlation values are calculated as follows: For each network we apply noise models at different noise levels (0% to 30%), and create ten networks for each case on which the results are averaged. For the varying noise levels, we compute the value of the network features (definitions of the non-standard features are provided in Table II) for each network in the set as well as the values of permanence, modularity, closeness centrality and PageRank listed earlier. We then compute the Pearson correlation coefficient of these values. The results are given in Fig. 4.

The values show that PageRank is not strongly correlated to any of the features at all. Modularity is strongly correlated to features that are defined by the community structure, such as intra and inter edges and their ratio. On the other hand both closeness centrality and permanence are strongly correlated with intrinsic network features such as diameter and average path length.

There are two interesting things to note. First, depending on the network, closeness centrality and permanence, can be either positively (LFR, $\mu = 0.1$) or negatively (football) correlated with the disassortative degree. Second, even though permanence is based on community assignment, it is not well correlated to any of the community dependent features. We believe that this is due to the local effect of permanence. Unlike modularity, it does not consider all edges within or outside a community, but only those to which it is tightly connected.

A question might arise that if both permanence and closeness centrality are strongly correlated to network features why does closeness not show a similar sensitivity. The answer lies in the range of values. In our experiments the highest and lowest values of permanence were (0.391 - 0.597) for LFR ($\mu = 0.1$) and (0.266 - 0.089) for football, whereas those for closeness centrality were (0.294 - 0.258) for LFR ($\mu = 0.1$) and (0.399 - 0.344) for football. Clearly, the difference is much more pronounced in the case of permanence.

Also note that closeness centrality is a more global metric as compared to permanence, in that it is a function of all the geodesics across the network. In contrast, permanence as discussed earlier is localized. Therefore, once the communities are known, permanence is faster to compute, of complexity $O(d_2)$, d_2 being the distance-2 degree, per vertex than closeness centrality, which is $O(V + E)$ per vertex.

V. Permanence based spreading

In this section we investigate the effectiveness of permanence in the context of message spreading.

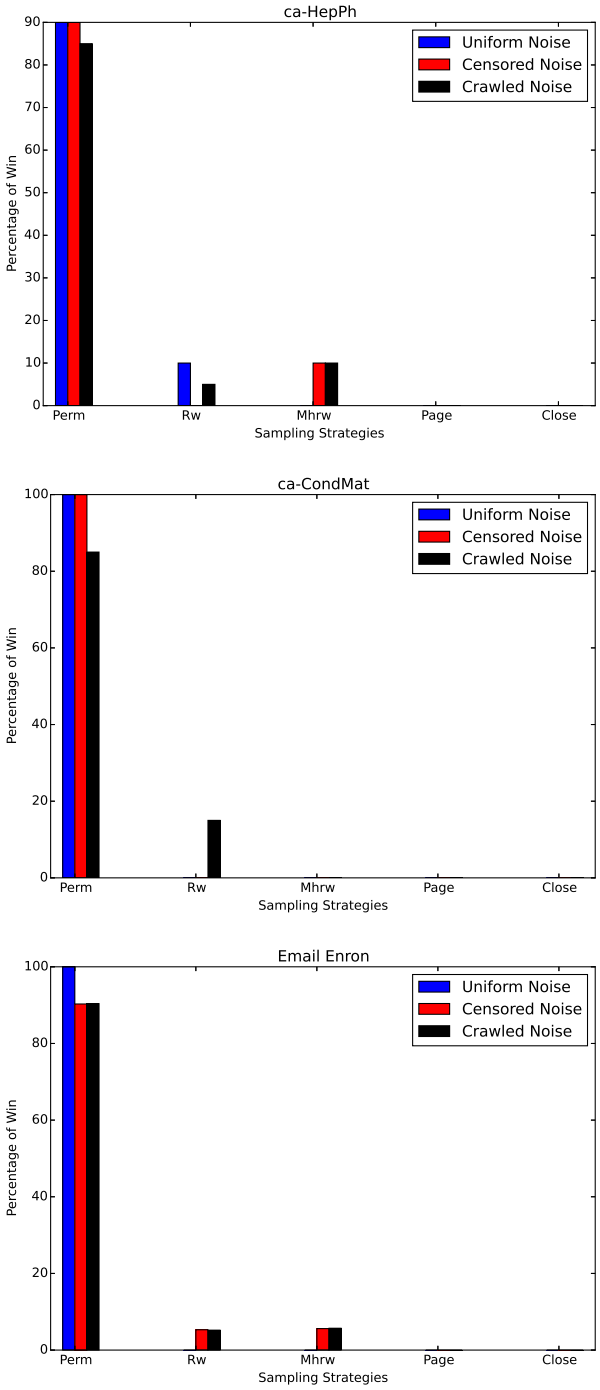

Fig. 6. Comparison of the percentage of cases in which a particular sampling method wins i.e., produces the least D-statistic of clustering co-efficient.

Methodoloy: As has been outlined in case of the baselines in Section III, we select initiator nodes from different communities preferentially based on their permanence values. For different levels of noise, we compute the time (in terms of the number of iterations) required to broadcast the message in the whole network and compare the values in case of permanence based initiator selection with that of the other baselines (i.e.,

Properties	Definition
Edge connectivity	Minimum edge connectivity (number of edges that must be removed in order to make the graph disconnected) over all pairs of vertices
Robustness	$\frac{\text{average sum of inverse geodesics between all pairs of nodes in the noisy network}}{\text{average sum of inverse geodesics in the original network}}$ [21], [22]
Inter edges	Edges interlinking communities expressed as a fraction of all edges
Intra edges	Edges internal to a community expressed as a fraction of all edges
Inter_intra	Fraction of inter over intra edges

TABLE II
DEFINITIONS OF THE NON-STANDARD NETWORK FEATURES.

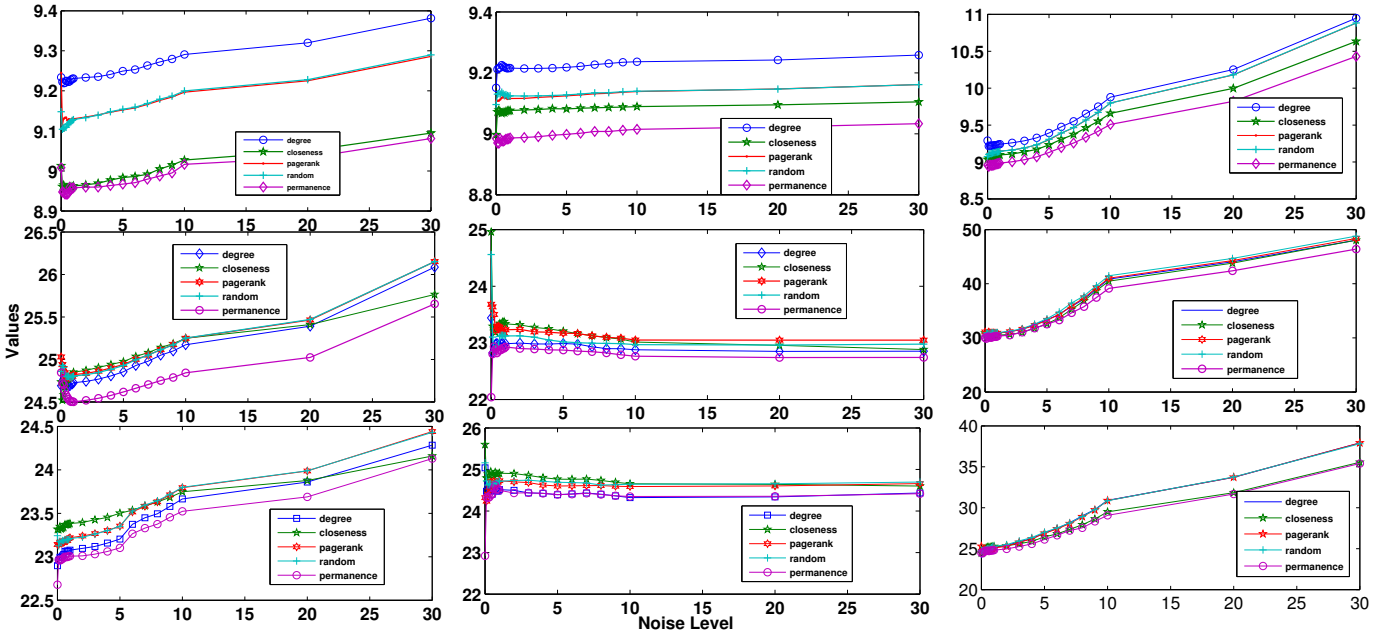

Fig. 5. Time required to broadcast a message for different initiator selection mechanisms (permanence, closeness, PageRank) and varying noise levels. The top panel is for the LFR network ($\mu = 0.1$), the middle is for the LFR network ($\mu = 0.3$) and the bottom panel is for the football network. The first column of figures represents results for uniform noise, the middle column represents results for censored noise and the last column represents results for the crawled noise.

random selection, degree based selection, closeness based selection and PageRank based selection.

Results: In Fig. 5, we plot the time required to broadcast for different levels of noise. For each noise level the results are averaged over 100 different runs. For all the noise models and all the datasets, permanence based initiator selection results in the least time required to broadcast the message. The main reason for this behaviour is that permanence is known to nicely arrange the nodes in a community in a core-periphery structure where the nodes with higher permanence are in the core of the community and those with lower permanence are in the periphery [23].

VI. SAMPLING FROM NOISY NETWORKS

In this section, we describe the sampling experiments and the associated results. Inspired by Maiya et al. [24], we mainly test how well the cluster structure of the sampled subnetwork corresponds to the original (noiseless) network.

Methodology: We obtain subnetworks from each network in the dataset and for different noise levels as already mentioned in Section III. The samples are obtained using the baseline algorithms discussed in Section III as well as preferentially based on permanence. In order to investigate how well the cluster structure is retained by the obtained samples we measure the following quantity – the D-statistic of clustering coefficient [25]. In specific, we compute the D-statistic[4] between the clustering coefficient distribution of the nodes in the original (noiseless) network and the nodes in the noisy sample. Note that the lower the D-statistic the better is the correspondence.

Results: From each original network and each noise model, we construct noisy versions at different noise levels varying from 0% to 20%. For each noisy version (at a particular noise level) we construct samples of different sizes having nodes varying from 5% to 20%. For each (noise level, sample size) combination we obtain the D-statistic of clustering coefficient. Since this is a stochastic process, for a particular noise level and a sample size, the results are averaged over ten different runs. In Fig. 6, we plot for the different datasets and for different noise models, in how many unique (noise level, sample size) pairs permanence based sampling obtains a lower D-statistic of clustering co-efficient compared to the other

[4]The D-statistic is defined as $D = max_x\{|F'(x) - F(x)|\}$ where x spans over the range of the random variables, and F and F' are the two empirical cumulative distribution functions of the data.

Networks	uniform noise	censored nose	crawled noise
Email Enron, (0.7156, 0.3320)	(**0.4888**, 0.3012)	(**0.5471**, 0.3300)	(**0.6434**, 0.3172)
	(0.4582, 0.2975)	(0.5032, 0.3271)	(0.6114, 0.3167)
	(0.4574, 0.2955)	(0.5158, 0.2955)	(0.6042, 0.3180)
ca-HepPh, (0.5989, 0.3714)	(**0.5339**, 0.3210)	(**0.5798**, 0.3412)	(**0.6511**, 0.3304)
	(0.4657, 0.3218)	(0.4867, 0.3471)	(0.5834, 0.3366)
	(0.4747, 0.3298)	(0.5088, 0.3504)	(0.6023, 0.3409)
ca-CondMat, (0.7054, 0.3333)	(**0.4356**, 0.3507)	(**0.5236**, 0.3553)	(**0.4158**, 0.3826)
	(0.3739, 0.3375)	(0.4682, 0.3499)	(0.3988, 0.3794)
	(0.3736, 0.3433)	(0.4754, 0.3543)	(0.3918, 0.3783)

TABLE III

TABLE SHOWING THE (MEAN CLUSTERING COEFFICIENT, STANDARD DEVIATION OF THE CLUSTERING COEFFICIENT) FOR THE THREE ORIGINAL NETWORKS AND FOR SUBNETWORKS WITH 20% SAMPLE SIZE SAMPLED FROM 20% NOISY VERSIONS OF THE ORIGINAL NETWORK. THE FIRST ENTRY IN EACH CELL IS FOR PERMANENCE BASED SAMPLING, THE SECOND ROW IS FOR THE RWR BASED SAMPLING AND THE THIRD ROW IS FOR THE MHRW BASED SAMPLING. THE BOLD ENTRIES INDICATE THE CLOSEST MATCH OF AVERAGE PERMANENCE WITH THE ORIGINAL NETWORK.

schemes. The results this figure clearly indicates that permanence based sampling by far outperforms all the other methods for all the networks and all the noise models considered, thus indicating that this is the case where the cluster structure is best preserved.

In order to make the observations further evident we report in Table III the mean and the standard deviation of the clustering coefficient of the three original networks and compare these values with subnetworks having sample size 20% and sampled from a 20% noisy version of the original network. The results are reported for permanence based sampling and two of its other immediate competitors – RWR based sampling and MHRW sampling. For all the three noise models and all the three networks, permanence outperforms both of its closest competitors. Note that these results are representative and holds true for all other sample sizes and all other low to moderate noise levels.

It is important to note that, one criticism to the use of permanence for sampling could be that the community information from the original network is required which might be difficult to compute and beat the whole purpose of sampling. However, note in many applications where network data needs to sampled from a larger (and possibly noisy) dataset, community information is already available; for instance, in an affiliation network the information related to the subscriptions of users to different interest groups (corresponding to communities) are usually known for the entire sample collected. In such cases, if only the local internal and external communities of the user are known, permanence can be directly computed without the necessity of the knowledge of the community structure of the whole network.

VII. PROPERTIES OF PERMANENCE

Our experimental results demonstrate that the mean permanence over the network is more sensitive to noise than other metrics (Section IV) and high permanence vertices are more effective in applications, such as spreading of messages and sampling of networks. This dual and seemingly opposite characteristic of being both sensitive and robust makes permanence a very useful metric indeed.

In order to understand the behavior of permanence, we perform the following experiments. We choose uniform noise and the LFR ($\mu = 0.1, 0.3$) and the football network as some of the representative cases for synthetic and real-world networks. The results are similar for other noise models as well as other networks.

A. Sensitivity of the factors comprising permanence

We investigate how different components of permanence change under uniform noise. Recall that permanence is computed as:
$P(v) = [\frac{I(v)}{E_{max}(v)} \times \frac{1}{D(v)} - (1 - c_{in}(v))]$
where $I(v)$ is the number of neighbors of v in its community, $E_{max}(v)$ is the maximum number of connections of v in its neighbouring community, $D(v)$ is the degree of v and $c_{in}(v)$ is the internal clustering coefficient of the internal nodes in the community of v.

We break the formula into two parts $PI = \frac{I(v)}{E_{max}(v)} \times \frac{1}{D(v)}$ and $c_{in}(v)$, and observe how they change over uniformly applied noise. The results in Fig. 7 show that PI remains relatively constant, whereas the internal clustering coefficient is the major contributor to the change in permanence. When we contrast this result with the main factors in modularity (Fig. 7), namely the internal and external edges, we see that each factor remains relatively constant.

B. Change in permanence profile for different noise models

Next we study how the permanence of the individual vertices are affected by noise. To do so, we divide the values of permanence, from -1 to 1, into 20 equal sized bins, such that bin 1 contains all vertices of permanence values form -1 to -0.9, and so on. We then plot in Fig. 8, how many vertices fall in each of the bins. In each case, the profile is shown for different noise levels starting from the original network. The permanence profile shifts toward lower numbered bins as the noise level increases. This is due to the fact, that when the community structure is good, there is a greater distribution of high permanence nodes that constitutes the base of good communities, but as noise increases there are more vertices with lower permanence.

C. Rank of high permanence vertices under noise

In our third experiment we compare permanence to closeness and PageRank to check how their top ranking vertices

alter under uniform noise. Here we first identify the top 10% of the high valued vertices for each centrality metric. Then for each noise level we compute the new top ranked vertices. We compute the Jaccard Index [26] between the original vertex set and the new one obtained from the noisy network. A high Jaccard index (maximum value 1) indicates that most of the top ranked vertices are retained under noise, and a low value (minimum 0) indicates that the set has changed completely.

As can be seen in Fig. 9, the Jaccard Index deteriorates much more slowly for permanence, whereas even the lowest level noise brings the Jaccard index to zero, for the other two centrality metrics in most of the cases. This indicates that the ids of the high valued permanence vertices remain relatively constant under noise.

These three experiments together highlight the unique properties of permanence and provides rationale of why it is effective both for evaluating noise and maintaining robustness of the results. The sensitivity experiments show that the internal clustering coefficient is most affected by noise. We posit that since uniform noise is likely to delete any edge with equal probability, on an average the ratio of the internal to external edges, or the degree does not alter (as also evidenced by our experiments). However, the change is more prominent within the cluster.

This effect is highlighted in the second experiment on permanence profile. Lower level vertices are more affected, and together they reduce the overall mean value of permanence making it sensitive to changes in the network. The third experiment related to ranking shows that even though the value decreases, the high permanence vertices, still retain their high ranks for low levels of noise. Thus, they still retain their relative centrality to the core of the community and thereby are effective for message broadcasting and sampling, even under noisy networks.

VIII. Related Work

Noise in networks has been studied very widely by network scientists since this might have both direct/indirect influence to the analysis and the predictions made thereof. In [27] the authors showed how noise could get introduced into the network due to incompleteness and inaccuracy of data collected from experiments. Kossinets [28] studied how noise in static networks affects the network centrality measures. In similar lines [29] outlined methods to measure centrality and cluster structures in networks with uncertain topologies. There have also been works to mine maximal cliques [30] from uncertain graphs as well as subgraph pattern matching [31] for uncertain graphs. [32] analyzed robustness of different network measures to different link errors. The effect of different forms of noise to community detection have been studied in [33]. Attack vulnerability have also been studied in presence of errors in networks [34]. Noise in networks can also arise from incomplete information due to privacy constraints as described in [35]. The entity resolution problem has also been tackled in the context of networks with limited information [36], [37]. A slightly different approach have been taken by Liu et al [38] where the authors combined for the first time the two problems – incomplete network discovery and community detection into a single framework. As an alternative approach, authors in [39] have tried to complete a network by developing algorithms to infer missing nodes and links in a network. There have also been works like [14], [40] that focused on the effect of various kind of synthetic noise effect on the centrality measures, community normalised mutual information and several other network measurements.

There have been a lot of works on message spreading and information diffusion; see [41] for a comprehensive review. There are broadly two approaches – explanatory models and predictive models. In the former category [42] attempts to investigate correlations in the infection times of nodes to infer the structure of the spreading cascade. In [43], the authors propose to model the diffusion process as a spatially discrete network. In a followup work [44] proposes a time-varying inference algorithm, to provide on-line estimates of the structure and temporal dynamics of a network. In the latter category most works attempt to predict how the diffusion process would unfold in the network at long times. These approaches are either graph based [45], [46] or non-graph based [47], [48].

Sampling of real-world networks is an equally challenging task that can be viewed as an inverse problem. There have been plethora of research in sampling trying to preserve the structural property of network in the representative sample following the seminal paper from Leskovec et al. [25]. There have been several works that preserve network properties like degree distribution, clustering coefficient, and so on [49], [50] using a suite of various stochastic procedures. Recently, Maiya et al. [24] showed the usefulness of sampling while considering the community structure of the network.

In this work, we combined these two objectives for the first time and proposed a centrality metric that can be shown to be appropriately sensitive to noise, efficient in spreading messages in noisy versions of a network as well as in accurate sampling of subnetworks from noisy networks.

IX. Conclusion

In this work, we have shown through rigorous experiments that permanence can be ideal quality metric for noisy incomplete networks. A key observation is that while the measure is appropriately sensitive to different noise levels, the high permanence nodes are almost unaffected by the application of noise thus making the measure at the same time very robust. Our findings can be summarised as:

- Permanence is the most appropriately sensitive metric among a wide range of other metrics tested in this work.
- Permanence is nicely correlated to various structural features of a network.
- Initiator selection based on permanence enables a faster message broadcast compared to other metrics. This observation holds even when the underlying network is noisy.
- Sampling based on permanence produces a cluster structure that corresponds significantly more with the original

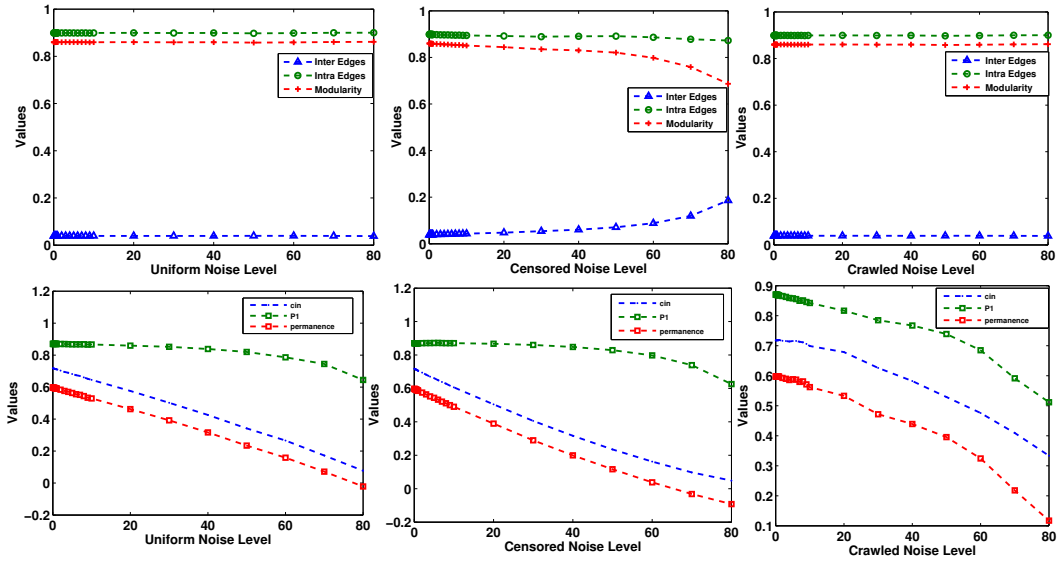

Fig. 7. The variation in the different components of permanence and modularity when the noise levels are varied. The top panel is for the LFR network ($\mu = 0.1$) and the bottom panel is for the football network.

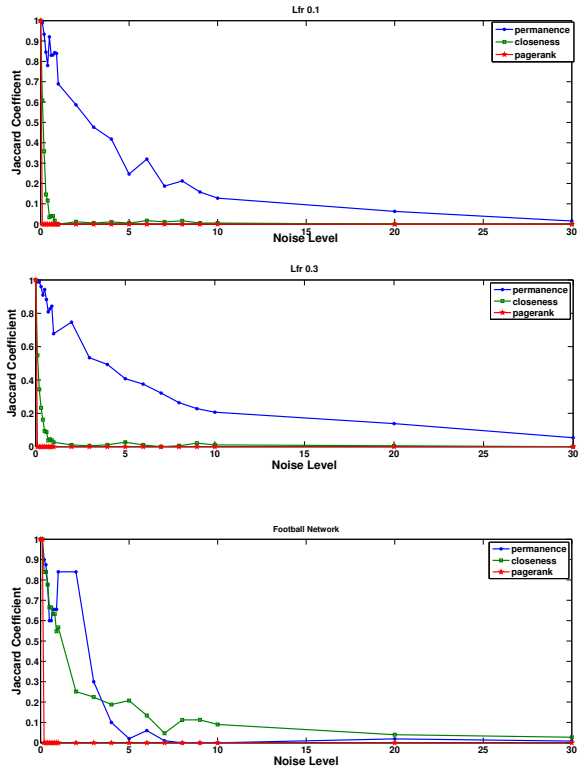

Fig. 9. The Jaccard Index between the top vertices of the original and the noisy networks for varying noise levels.

network than other baseline methods. Once again this observation extends to even the noisy samples.

There are quite a few interesting future directions of this study. First, we would like to perform a systematic analytical study to identify the differences between simple clustering coefficient of a node and its permanence. Second, we would like to investigate the analytical reasons for the stability of high permanence nodes and, thereby, propose an algorithm to automatically identify the level of noise up to which this stability persists. Third, we would like to perform a full-scale study of the effect of larger levels of noise on community detection algorithms and if permanence could be meaningfully used to obtain more accurate results for such extreme cases. Finally, we would like to propose a permanence based noise tolerant community detection algorithm that can on-the-fly perform network discovery and clustering at the same time.

ACKNOWLEDGMENT

SS and AM would like to acknowledge the financial support from the ITRA DISARM project from DeiTY.

REFERENCES

[1] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[2] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, and A.-L. Barabási, "The large-scale organization of metabolic networks," *Nature*, vol. 407, no. 6804, pp. 651–654, 2000.

[3] S. Milgram, "The small world problem," *Psychology today*, vol. 2, no. 1, pp. 60–67, 1967.

[4] D. J. Watts and S. H. Strogatz, "Collective dynamics of small-worldnetworks," *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[5] 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.

[6] S. Eubank, H. Guclu, V. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modelling disease outbreaks in realistic urban social networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[7] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical review letters*, vol. 86, no. 14, p. 3200, 2001.

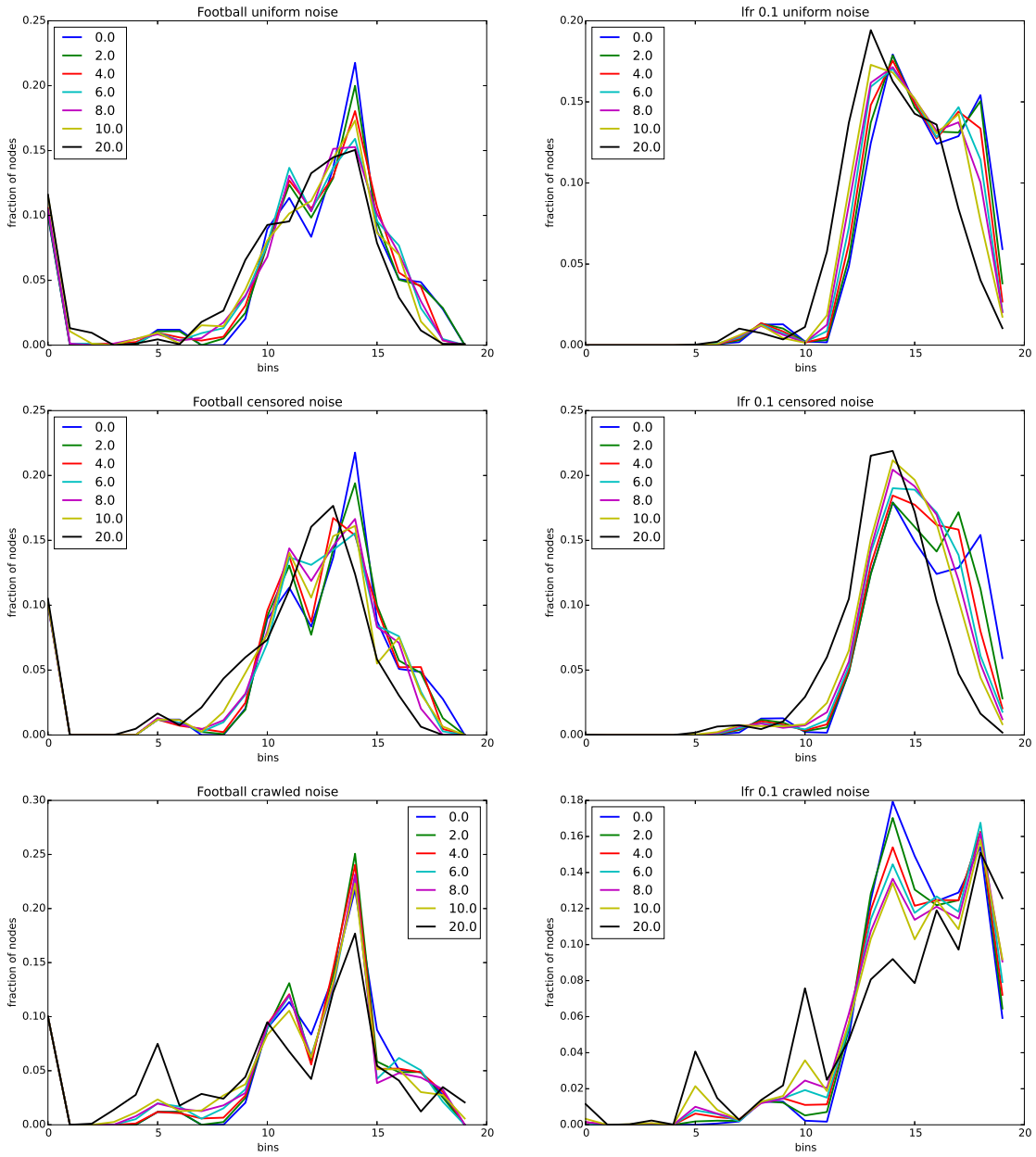

Fig. 8. Normalised permanence profile distribution for different noise levels. The results are shown for two different networks and the three different noise models.

[8] T. Chakraborty, S. Srinivasan, N. Ganguly, A. Mukherjee, and S. Bhowmick, "On the permanence of vertices in network communities," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2014, pp. 1396–1405.

[9] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the national academy of sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[10] A. Lancichinetti and S. Fortunato, "Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities," *Physical Review E*, vol. 80, no. 1, p. 016118, 2009.

[11] J. Leskovec, J. Kleinberg, and C. Faloutsos, "Graph evolution: Densification and shrinking diameters," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 1, no. 1, p. 2, 2007.

[12] M. Rosvall, D. Axelsson, and C. T. Bergstrom, "The map equation," *The European Physical Journal Special Topics*, vol. 178, no. 1, pp. 13–23, 2010.

[13] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney, "Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters," *Internet Mathematics*, vol. 6, no. 1, pp. 29–123, 2009.

[14] B. Yan and S. Gregory, "Finding missing edges and communities in incomplete networks," *J Phys A*, vol. 44, p. 495102, 2011.

[15] S. A. Catanese, P. De Meo, E. Ferrara, G. Fiumara, and A. Provetti, "Crawling facebook for social network analysis purposes," in *Proceedings of the international conference on web intelligence, mining and semantics*. ACM, 2011, p. 52.

[16] 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.

[17] E. O. Laumann, P. V. Marsden, and D. Prensky, "The boundary specification problem in network analysis," *Research methods in social network analysis*, vol. 61, p. 87, 1989.

[18] Y.-X. Zhu, L. Lü, Q.-M. Zhang, and T. Zhou, "Uncovering missing links with cold ends," *Physica A: Statistical Mechanics and its Applications*, vol. 391, no. 22, pp. 5769–5778, 2012.

[19] F. Chierichetti, S. Lattanzi, and A. Panconesi, "Rumour spreading and graph conductance." in *SODA*. SIAM, 2010, pp. 1657–1663.

[20] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "A walk in facebook: Uniform sampling of users in online social networks," *arXiv preprint arXiv:0906.0060*, 2009.

[21] S. Sur, N. Ganguly, and A. Mukherjee, "Attack tolerance of correlated time-varying social networks with well-defined communities," *Physica A: Statistical Mechanics and its Applications*, vol. 420, pp. 98–107, 2015.

[22] S. Scellato, I. Leontiadis, C. Mascolo, P. Basu, and M. Zafer, "Understanding robustness of mobile networks through temporal network measures," in *INFOCOM, 2011 Proceedings IEEE*. IEEE, 2011, pp. 1–5.

[23] T. Chakraborty, S. Srinivasan, N. Ganguly, A. Mukherjee, and S. Bhowmick, "Permanence and community structure in complex networks." ACM, submitted.

[24] A. S. Maiya and T. Y. Berger-Wolf, "Benefits of bias: Towards better characterization of network sampling," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2011, pp. 105–113.

[25] 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.

[26] J. C. Gower *et al.*, "Measures of similarity, dissimilarity and distance," *Encyclopedia of statistical sciences*, vol. 5, no. 397-405, p. 3, 1985.

[27] R. Guimerà and M. Sales-Pardo, "Missing and spurious interactions and the reconstruction of complex networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 52, pp. 22 073–22 078, 2009.

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

[29] J. J. Pfeiffer III and J. Neville, "Methods to determine node centrality and clustering in graphs with uncertain structure," *arXiv preprint arXiv:1104.0319*, 2011.

[30] A. P. Mukherjee, P. Xu, and S. Tirthapura, "Mining maximal cliques from an uncertain graph," *arXiv preprint arXiv:1310.6780*, 2013.

[31] N. Vesdapunt and H. Garcia-Molina, "Identifying users in social networks with limited information," 2014.

[32] J. Platig, E. Ott, and M. Girvan, "Robustness of network measures to link errors," *Physical Review E*, vol. 88, no. 6, p. 062812, 2013.

[33] L. Wang, J. Wang, Y. Bi, W. Wu, W. Xu, and B. Lian, "Noise-tolerance community detection and evolution in dynamic social networks," *Journal of Combinatorial Optimization*, vol. 28, no. 3, pp. 600–612, 2014.

[34] L. B. Booker, "The effects of observation errors on the attack vulnerability of complex networks," DTIC Document, Tech. Rep., 2012.

[35] F. Chierichetti, A. Epasto, R. Kumar, S. Lattanzi, and V. Mirrokni, "Efficient algorithms for public-private social networks," in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2015, pp. 139–148.

[36] W. E. Moustafa, A. Kimmig, A. Deshpande, and L. Getoor, "Subgraph pattern matching over uncertain graphs with identity linkage uncertainty," in *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*. IEEE, 2014, pp. 904–915.

[37] V. Verroios and H. Garcia-Molina, "Entity resolution with crowd errors," 2015.

[38] J. Liu, C. Aggarwal, and J. Han, "On integrating network and community discovery," in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*. ACM, 2015, pp. 117–126.

[39] M. Kim and J. Leskovec, "The network completion problem: Inferring missing nodes and edges in networks." in *SDM*. SIAM, 2011, pp. 47–58.

[40] 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.

[41] A. Guille, H. Hacid, C. Favre, and D. A. Zighed, "Information diffusion in online social networks: A survey," *SIGMOD Rec.*, vol. 42, no. 2, pp. 17–28, 2013.

[42] M. Gomez Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: ACM, 2010, pp. 1019–1028.

[43] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proceedings of the 28th International Conference on Machine Learning, ICML 2011, Bellevue, Washington, USA, June 28 - July 2, 2011*, 2011, pp. 561–568.

[44] M. Gomez Rodriguez, J. Leskovec, and B. Schölkopf, "Structure and dynamics of information pathways in online media," in *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, ser. WSDM '13. New York, NY, USA: ACM, 2013, pp. 23–32.

[45] M. Granovetter, "Threshold models of collective behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, 1978.

[46] A. Guille and H. Hacid, "A predictive model for the temporal dynamics of information diffusion in online social networks," in *Proceedings of the 21st International Conference on World Wide Web*, ser. WWW '12 Companion. New York, NY, USA: ACM, 2012, pp. 1145–1152.

[47] J. Leskovec, M. McGlohon, C. Faloutsos, N. S. Glance, and M. Hurst, "Patterns of cascading behavior in large blog graphs," in *Proceedings of the Seventh SIAM International Conference on Data Mining, April 26-28, 2007, Minneapolis, Minnesota, USA*, 2007, pp. 551–556.

[48] J. Yang and J. Leskovec, "Modeling information diffusion in implicit networks," in *Proceedings of the 2010 IEEE International Conference on Data Mining*, ser. ICDM '10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 599–608.

[49] M. Gjoka, M. Kurant, C. T. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *INFOCOM, 2010 Proceedings IEEE*. IEEE, 2010, pp. 1–9.

[50] A. Rezvanian and M. R. Meybodi, "Sampling social networks using shortest paths," *Physica A: Statistical Mechanics and its Applications*, vol. 424, pp. 254–268, 2015.