Measuring the Sampling Robustness of Complex Networks

Katchaguy Areekijseree Sucheta Soundarajan
{kareekij,susounda}@syr.edu

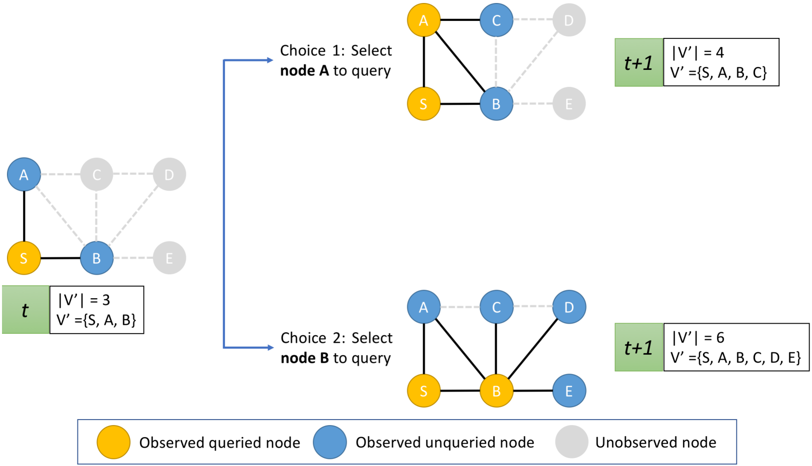
Syracuse University, Syracuse, NY, USA

ASONAM 2019

Introduction I

- Many researchers are interested in complex networks.
- There are many ways to collect network data (network sampling).
  - e.g. API, pen-and-paper questionnaires, surveys, interviews.
  - When query, a list of neighboring nodes is returned in response.
- However, *errors* may occur during data collection process.
  - e.g. mistakes from participants' answer, bug in a web crawler, adversary.
- Errors may lead to errors in a subsequent network analysis.
- Thus, it is important for a data analyst to know if a collected sample is trustworthy.

Data Collection



In this work

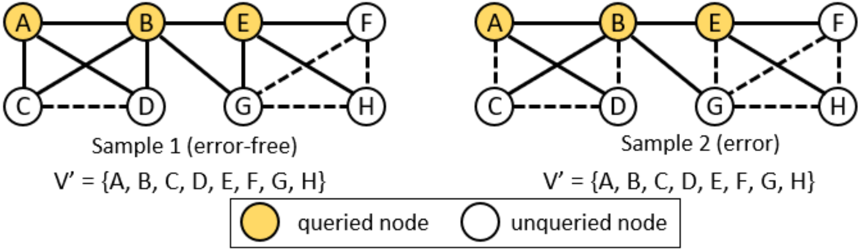
1. We introduce a new robustness measure called "Sampling Robustness".
2. We model error as random edge deletion.
3. **Goal**: Investigate how sampling robustness of a network changes due to random edge deletion.
4. Sampling robustness is highly correlated with properties of the network and obtained sample.
5. We present regression models for estimating sampling robustness.

What is Sampling Robustness?

If a network G has **high sampling robustness**, the performance of a crawler C on G will be consistent regardless of the existence of the errors.

Random Error

- We consider *random edge deletion*.
- Each query, some fraction of edges is missing.
- Each returned edge has a probability p that it will be removed.



Sampling Robustness I

Sampling Robustness

$$R_p(G, C) = \frac{sim(M(S), M(S^{'}))}{\bar{R}_0}$$

where

- S represents the *complete* sample.
- $S^{'}$ represents the sample obtained by the crawler C with errors.
- $M(\cdot)$ is an application-specific function which characterizes the performance of the crawler C when it generates a sample.
- $sim(\cdot, \cdot)$ is a similarity measure.

Sampling Robustness II

Performance Measure
In this work, we use $M(\cdot)$ as a size of the sample (node coverage).

$$M(S) = |\{v \in V_s^{'}, V_s^{'} \subseteq V\}|$$

Similarity Measure
Thus, the similarity of S and $S^{'}$ can be computed using Canberra distance.

$$sim(M(S), M(S^{'}) = 1 - d_{canberra}(|V_s^{'}|, |V_{s'}^{'}|)$$

Other performance and similarity measures

Sampling Robustness can be calculated by using other performance measures.

Performance Measure, $M(\cdot)$	Type	Similarity measure, $sim(S, S')$
number of nodes or edges found	number	$1 - d_{canberra/L_1/L_2}$
distinct nodes in the sample	a set	Jaccard similarity
communities membership	a set of set	NMI, Partition similarity
degree distribution of the sample	distribution	1-KS statistic

Network Crawling Technique

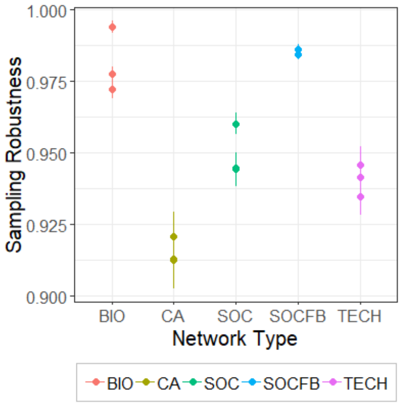
We consider three popular crawling algorithms

1. **Breadth-first search (BFS)**: a crawler selects the node that has been in the list of unqueried nodes the longest (FIFO).
2. **Random walk (RW)**: the crawler transitions to a neighbor of the node that was just queried at random.
3. **Maximum observed degree (MOD)**: This crawler selects the unqueried node with the highest degree

Sampling Robustness and Network Type

Type	Network	$\|V\|$	$\|E\|$	$\bar{d}$	$\bar{cc}$	λ_1
CA	Erdos992	4991	7428	2.977	0.08352	15.13
	HepTh	8638	24806	5.743	0.4816	31.03
	GrQc	4158	13422	6.456	0.5569	45.62
BIO	CE-GN	2215	53680	48.47	0.1843	96.22
	CE-PG	1692	47309	55.92	0.4467	152.6
	SC-GT	1708	33982	39.79	0.3491	109.9
SOCFB	Amherst41	2235	90954	81.39	0.3104	137.1
	Colgate88	3482	155043	89.05	0.2673	141.9
	Bowdoin47	2250	84386	75.01	0.289	124.2
SOC	Hamsterster	2000	16097	16.1	0.54	50.02
	Advogato	5054	39374	15.58	0.2526	70.51
	Wiki-Elec	7066	100727	28.51	0.1418	138.1
Tech	PGP	10680	24316	4.554	0.2659	42.44
	Router-RF	2113	6632	6.277	0.2464	27.67
	WhoIS	7476	56943	15.23	0.4889	150.9

Table: Statistics of network.



Observation

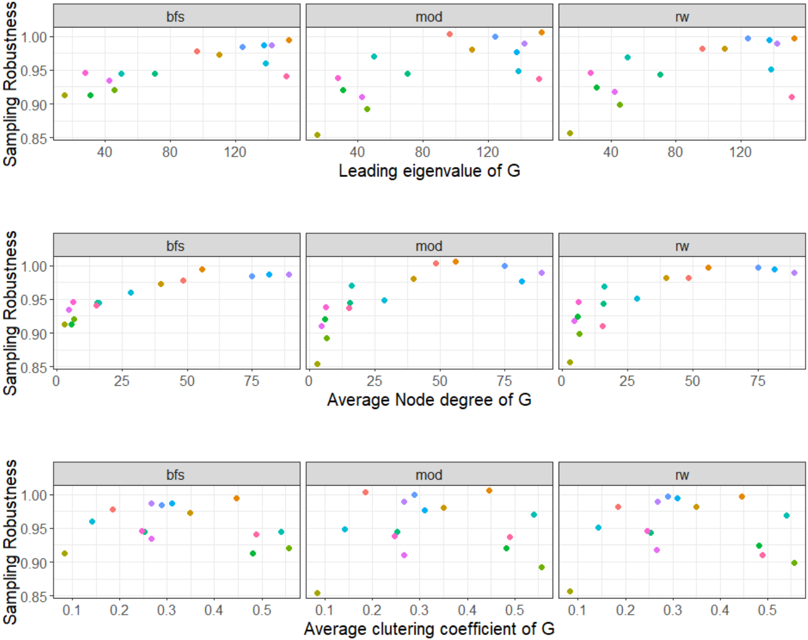
Different level sampling robustness on different network types.

Characterizing Sampling Robustness

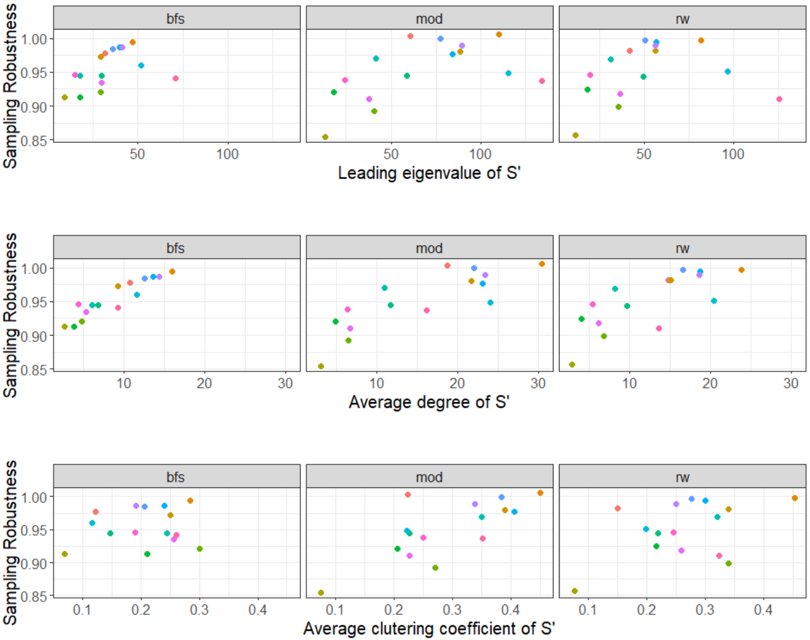
We investigate 3 properties that we believe support a crawler in expanding sample's boundary.

1. The largest eigen value of adjacency matrix A.
   - Forecasting epidemic spreading process (e.g. SIR model).
   - Epidemic threshold $\tau = \frac{1}{\lambda_1}$.
2. Average degree (average number of neighboring nodes)
   - A crawler quickly expands the sample when average degree is large.
3. Average clustering coefficient.
   - It measures how well nodes are connected.
   - Intuitively, when nodes are densely connected, the crawler discovers nodes quickly.

Robustness vs Properties of an Original Network G



Robustness vs Properties of a collected sample S'



How we can compute Sampling Robustness ?

Sampling Robustness

$$R_p(G, C) = \frac{sim(M(S), M(S^{'}))}{\bar{R}_0}$$

In order to calculate sampling robustness, we need an information about complete sample S.
However, we only obtain only $S^{'}$ in real scenario.

Estimating Sampling Robustness I

Given an obtained sample S, we present a linear regression model for estimating a sampling robustness of any network.

$$\hat{R}_p = \beta_1 p + \beta_2 \bar{d}^{'} + \beta_3 \lambda_1^{'} + \beta_4 \bar{cc}^{'} + \beta_5 (cc^{'} \times \bar{d}^{'}) + b,$$

where

- $\beta_1 ... \beta_k$ are the coefficients.
- $\bar{d}^{'}$ - average degree of S.
- $\bar{cc}^{'}$ - average clustering coefficient of S.
- $\lambda^{'}$ - leading eigenvalue of S.
- p - error probability

Estimating Sampling Robustness II

$$\hat{R}_p = \beta_1 p + \beta_2 \bar{d}' + \beta_3 \lambda_1' + \beta_4 \bar{cc}' + \beta_5 (cc' \times \bar{d}') + b,$$

Estimating error probability p:

- Perform multiple queries on the same node, k.
- Counting the number of times a particular edge is duplicated, k_e.
- Thus, $p = 1 - \frac{k_e}{k}$.

Estimating Sampling Robustness III

Model Training: We train our model from several sampled networks. In total, there are 2,200 samples.

Table: Coefficients and intercept of each model

Model	β_1	β_2	β_3	β_4	β_5	b
M1-RW	-0.1843	0.0127	-0.0009	0.4374	-0.0245	0.8661
M2-BFS	-0.1951	0.0119	-0.0006	0.2165	-0.0250	0.9313
M3-MOD	-0.2199	0.0094	-0.0006	0.3928	-0.0152	0.8801

Model Evaluation

We generate several samples from these networks using BFS, MOD and Random walk crawler (600 samples).

Table: Statistics of network used for model testing.

Network	$\|V\|$	$\|E\|$	$\bar{d}$	$\bar{cc}$	λ_1
Hamilton46	2312	96393	83.38	0.2983	135.93
Trinity100	2613	111996	85.72	0.2903	135.83
Epinion	26588	100120	7.53	0.1351	66.206
Caida2007	26475	53381	4.03	0.2082	69.643

Model Evaluation - Results

	M1	M2	M3
MSE	0.00127	0.00089	0.00142
R^2	0.7258	0.7147	0.7440

Overall, our proposed models are capable of estimating the sampling robustness of a network G from a sample $S^{'}$ with very small MSE (< 0.0015) and a R-squared of up to 0.75.

Conclusion

- We present a novel network robustness measure called "sampling robustness".
- We demonstrate that each network types have different level of robustness.
- Sampling robustness is highly depends on an original network properties and also the properties of the obtained samples.
- We can estimate the robustness from these properties.

Thank You

Questions?

`kareekij@syr.edu`