Sensitivity of Community Structure to Network Uncertainty

Fragkiskos D. Malliaros | UC San Diego

Marc Mitri
Michalis Vazirgiannis | École Polytechnique

SIAM International Conference on Data Mining (SDM)
Houston, Texas
April 27-29, 2017

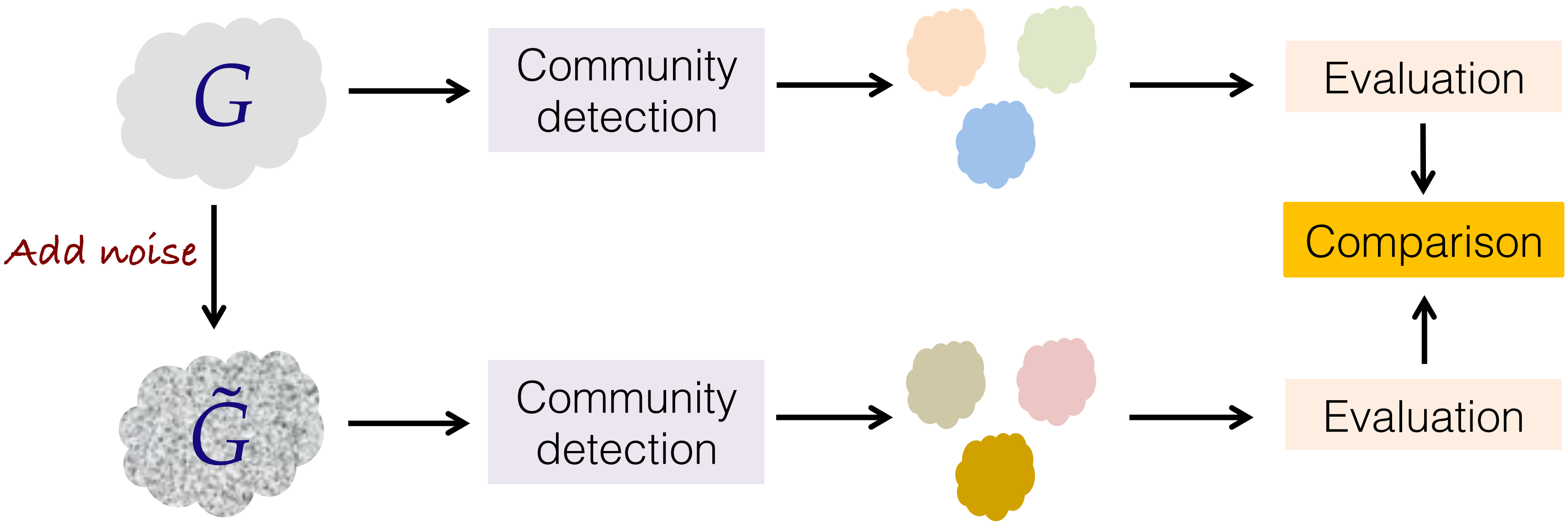
Introduction and Motivation

- Real networks are often noisy and incomplete
  - Noise introduced during the data collection process
  - Uncertainty due to privacy preserving reasons
  - …

- Motivation: How robust (i.e., stable) are the results of a community detection algorithm under network uncertainty?
  - How do we define network uncertainty?
  - Model uncertainty as a graph perturbation process

- Our goal: study the behavior of community detection algorithms under several graph perturbation strategies

Overview of Our Approach



Outline

- Graph perturbation strategies
- Community evaluation
  - Functional sensitivity
  - Structural sensitivity
- Experimental results
- Conclusions

Graph perturbation strategies

How to Model Uncertainty?

- Let G be the original graph and $\mathbb{G}(n)$ be a random graph model
- Then, the noise model $\theta(G, \mathbb{G}, \varepsilon_a, \varepsilon_d)$ using the random graph $\mathbb{G}(n)$ gives the probability of adding/deleting an edge (u, v) by

$$\mathbb{P}_\theta((u,v)) = \begin{cases} \varepsilon_a \mathbb{P}_{\mathbb{G}}((u,v)), & \text{if } (u,v) \notin E_G \\ \varepsilon_d \mathbb{P}_{\mathbb{G}}((u,v)), & \text{if } (u,v) \in E_G \end{cases}$$

Probabilities of edge addition and deletion

Probability of selecting edge (u, v)

- By XOR-ing the original graph with one realization $R \in \theta(G, \mathbb{G}, \varepsilon_a, \varepsilon_d)$ of the noise model, we obtain the perturbed graph $\tilde{G} = G \oplus R$

ERP Model

- Uniform perturbation model
  - $\mathbb{G} = \mathcal{G}(n, 1/n)$ is the Erdös-Rényi random graph model

- In this case, $\mathbb{P}_{\mathbb{G}}((u, v)) = 1/n$

Edges are added/removed independently

- Noise model:

$$ERP(G, \varepsilon_a, \varepsilon_d) = \theta(G, \mathcal{G}(n, 1/n), \varepsilon_a, \varepsilon_d)$$

Example

$$ERP(G, \underset{\substack{\downarrow \\ 10/n}}{\overset{\text{add}}{10}}, \underset{\substack{\downarrow \\ 20/n}}{\overset{\text{del}}{20}})$$

CLP Model

- Preferential perturbation based on the Chung-Lu random graph model

$$\mathbb{P}_{\mathbb{G}}((u,v)) \propto \kappa_u \cdot \kappa_v$$

Edges are added/removed with probability proportional to the degree of the endpoints

ConfMP Model

degree sequence

$\vec{\kappa} = \{\kappa_u\}$

- Configuration model $\mathbb{G} = \mathcal{G}(n, \vec{\kappa})$
- The number of edges is the same as in the original network
- Rewire a certain amount of edges under the constraint that $\vec{\kappa} = \{\kappa_u\}$ will remain the same after the perturbation

Probability of an edge between *u* and *v*

$$e_{uv} = 2mp_u p_v = 2m\frac{\kappa_u \kappa_v}{4m^2} = \frac{\kappa_u \kappa_v}{2m}$$

How do we measure sensitivity

Sensitivity of Community Structure

- Functional sensitivity

  How similar are the communities of the perturbed and unperturbed (original) graph?

- Structural sensitivity

  How do the structural properties of the communities change?

Functional Sensitivity

- Normalized Mutual Information (NMI)
  - 'NMI=0': independent communities
  - 'NMI=1': identical communities

$$I_{norm}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}$$

Higher value is better

- Variation of Information (VI)
  - 'VI=0': identical communities
  - 'VI=log(n)': maximum value

$$VI(X, Y) = H(X|Y) + H(Y|X)$$

Lower value is better

- Adjusted Rand Index (ARI)
  (based on counting of pairs of elements)
  - 'ARI=0': independent communities
  - 'ARI=1': identical communities

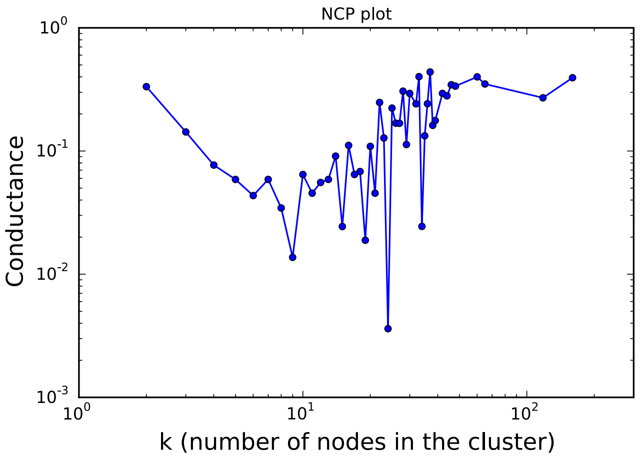
$$ARI(X, Y) = \frac{\overbrace{a + b}^{\text{\# of agreements}}}{a + b + \underbrace{c + d}_{\text{\# of disagreements}}}$$

Higher value is better

Structural Sensitivity

- Conductance $\phi(S) = \dfrac{\sum_{u \in S, v \notin S} A_{uv}}{\kappa_S}$ Lower value is better

- Network Community Profile Plot (NCP)



- Spectral Lower Bound λ_G
  - Algebraic connectivity: second smallest eigenvalue of the Laplacian matrix

Experimental Results

Community Detection Algorithms

- Fast greedy modularity optimization (FastGreedyMM) [Clauset at al '04]
- Louvain modularity optimization [Blondel et al. '08]
- Leading eigenvector [Newman '06]
- Spectral clustering [Ng et al. '02]
- Label propagation [Raghavan et al. '07]
- Metis [Karypis and Kumar '99]
- Infomap [Rosvall and Bergstrom '07]
- Walktrap [Pons and Latapy '05]

$$Q = \frac{1}{2m} \sum_{u,v} \left[A_{uv} - \frac{\kappa_u \kappa_v}{2m} \right] \delta(c_u, c_v)$$

The algorithms are publicly available (e.g., igraph library)

Datasets

Network	# of nodes	# of edges
AS-CAIDA	16,301	65,910
WIKI-VOTE	7,115	103,689
CA-GR-QC	5,242	14,496
CA-HEP-TH	9,877	25,998
P2P-GNUTELLA	6,301	20,777

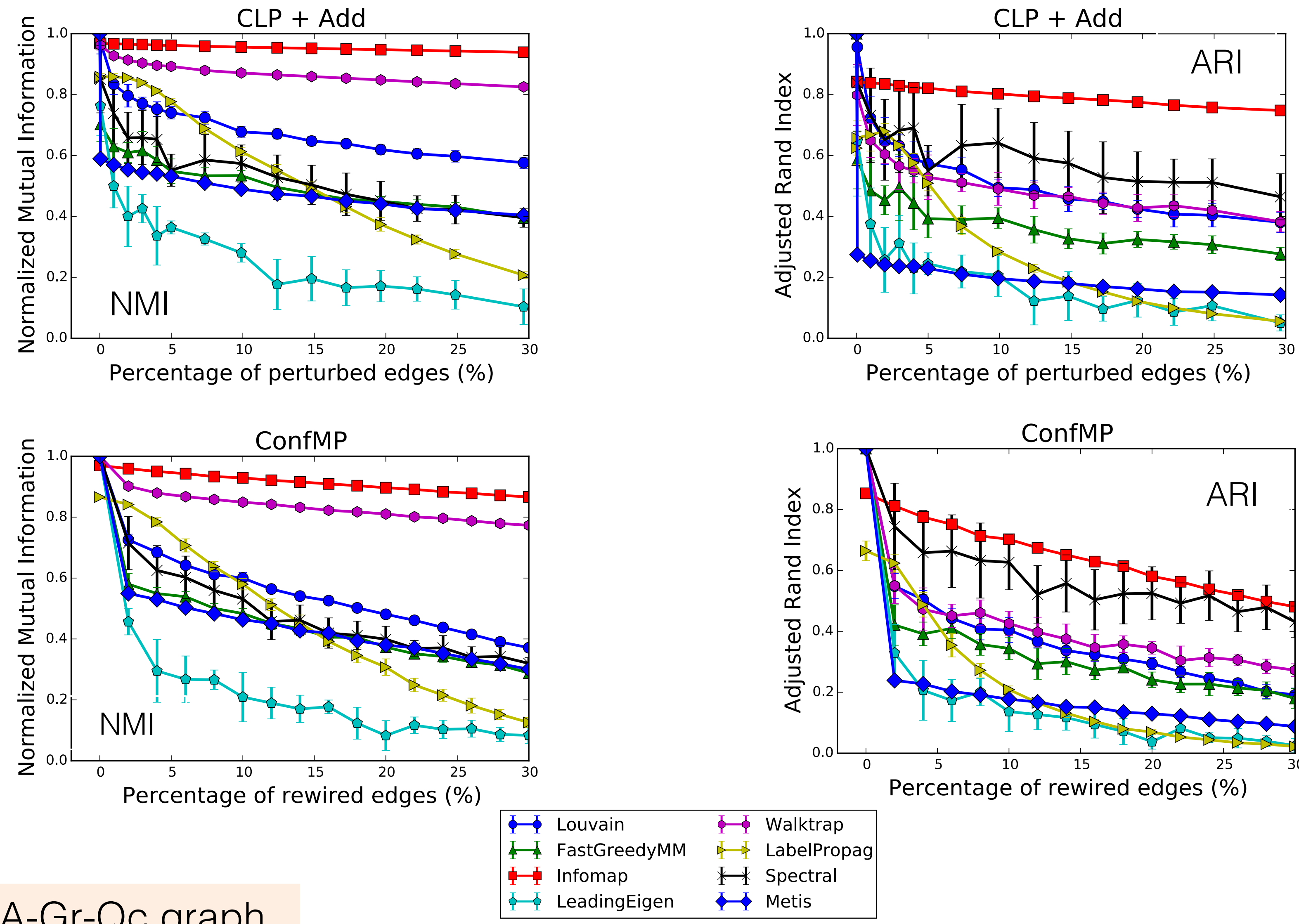
Source: http://snap.stanford.edu

Experimental Setup

- Graphs are unweighted and undirected (keep GCC only)
- The number of clusters for Metis and Spectral is set to be equal to the number of communities detected by Louvain algorithm (modularity optimization)
- Infomap and LabelPropag are not deterministic
  - Average over multiple runs for each noise level
- Examine various noise levels from 0% to 30%
  - Ensure that the perturbed graphs are still connected

Functional Sensitivity Analysis

How similar are the communities of the perturbed and unperturbed graphs?



CA-Gr-Qc graph

Observations

Observation

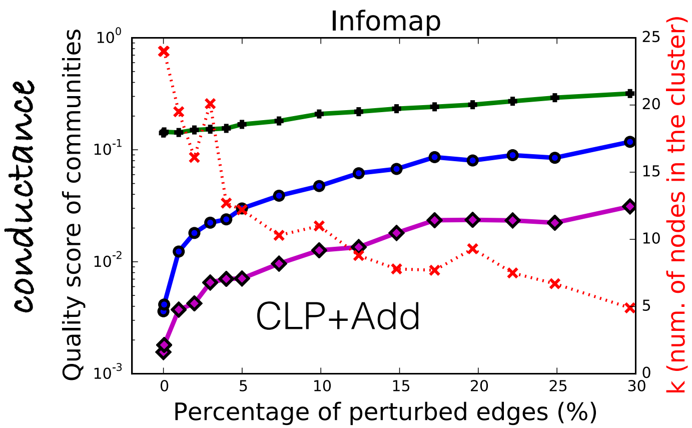
- Infomap is the most robust algorithm in almost all cases
  - High NMI and ARI values even for high perturbation levels
  - The output of the algorithm is stable
  - The Walktrap algorithm also performs very well

Stability of random walk based algorithms

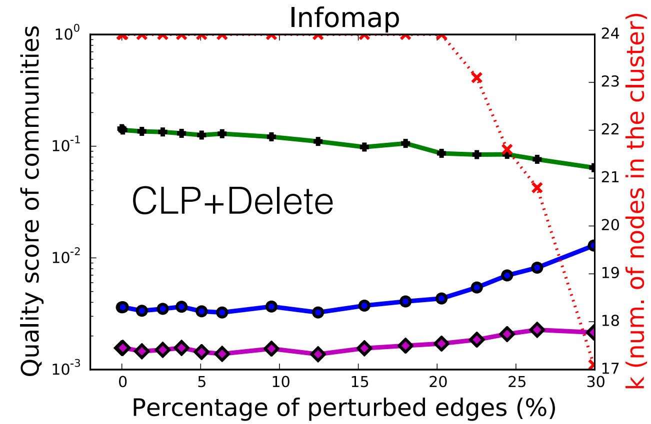
- Random walk based methods tend to be very robust to noise
  - Why? Stability of the eigenvectors of the transition matrix P of the random walk under perturbation

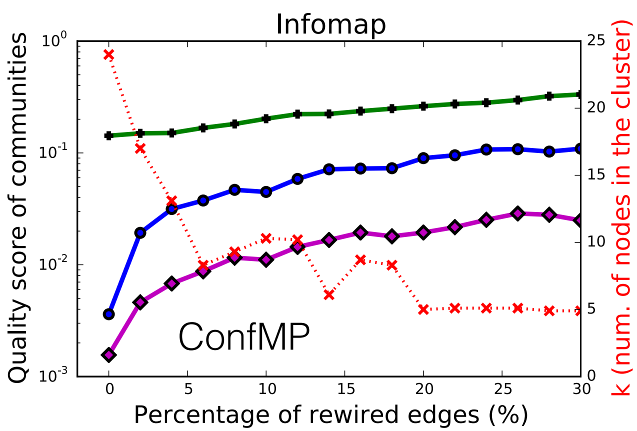
Structural Sensitivity Analysis

How do the structural properties of the communities change?



Size of the best cluster





Global min (of conductance scores
Median over all communities)
Spectral lower bound

(property of the data)

Observations

- Correlation between conductance (*real behavior*) and spectral lower bound (*theory*)
- Uptrend in CLP+Add and ConfMP
  - The quality of communities is reduced
  - Different behavior in edge deletions

CA-Gr-Qc graph

Related Work

- Few papers on the robustness of community detection algorithms
  - Mainly focus on the properties of spectral clustering
  - Robustness of spectral modularity optimization under the ConfMP model [Karrer, Levina and Newman '08]
  - Robustness w.r.t. the identification of ground truth communities [Yang and Leskovec '15]
  - Comparison of community detection algorithms based on artificial networks [Danon et al. '05], [Lancichinetti and Fortunato '09]

- Sensitivity analysis in other graph mining tasks
  - Web ranking algorithms [Ng et al. '02]
  - Influence maximization models [Adiga et al. '14]
  - Core decomposition [Adiga and Vullikanti '13]
  - Entity selection tasks (e.g., influence maximization) [Misra, Golshan and Terzi '12]

Conclusions

- Sensitivity of community structure under uncertainty
  – Functional and structural sensitivity analysis
  – Random walk based algorithms tend to be robust against noise
- Take home message: sensitivity as an additional evaluation tool for community detection algorithms
- Future work
  – More generalized theoretical analysis (beyond spectral and random walk based algorithms)
  – Sensitivity of local community detection algorithms

Thank You!

Project Website: fragkiskos.me/projects/communities_sensitivity