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
• 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

- Add noise

$G$ → Community detection

$\tilde{G}$ → Community detection

Evaluation → Comparison → Evaluation
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 $G(n)$ be a random graph model

• Then, the noise model $\theta(G, G, \varepsilon_a, \varepsilon_d)$ using the random graph $G(n)$ gives the probability of adding/deleting an edge $(u, v)$ by

$$P_\theta((u, v)) = \begin{cases} 
\varepsilon_a P_G((u, v)), & \text{if } (u, v) \notin E_G \\
\varepsilon_d P_G((u, v)), & \text{if } (u, v) \in E_G 
\end{cases}$$

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

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

- In this case, $\mathbb{P}_G((u,v)) = \frac{1}{n}$

- Noise model:

\[
\text{ERP}(G, \varepsilon_a, \varepsilon_d) = \theta(G, \mathcal{G}(n, \frac{1}{n}), \varepsilon_a, \varepsilon_d)
\]

Example

\[
\text{ERP}(G, 10, 20) \quad \frac{10}{n} \quad \frac{20}{n}
\]
• Preferential perturbation based on the Chung-Lu random graph model

\[ P_G((u, v)) \propto \kappa_u \cdot \kappa_v \]

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

- **Configuration model** \( G = G(n, \vec{k}) \)

- The number of edges is the same as in the original network

- Rewire a certain amount of edges under the constraint that \( \vec{k} = \{ \kappa_u \} \) 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

- **Variation of Information (VI)**
  - ‘VI=0’: identical communities
  - ‘VI=\log(n)’: maximum value

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

\[
I_{\text{norm}}(X, Y) = \frac{2I(X, Y)}{H(X) + H(Y)}
\]
Higher value is better

\[
VI(X, Y) = H(X|Y) + H(Y|X)
\]
Lower value is better

\[
ARI(X, Y) = \frac{a + b}{a + b + c + d}
\]
Higher value is better
Structural Sensitivity

- **Conductance**
  \[ \phi(S) = \frac{\sum_{u \in S, v \notin S} A_{uv}}{\kappa_S} \]
  Lower value is better

- **Network Community Profile Plot (NCP)**

- **Spectral Lower Bound** \( \lambda_G \)
  - **Algebraic connectivity**: second smallest eigenvalue of the Laplacian matrix
Experimental Results
Community Detection Algorithms

- Fast greedy modularity optimization (FastGreedyMM) [Clauset et 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

<table>
<thead>
<tr>
<th>Network</th>
<th># of nodes</th>
<th># of edges</th>
</tr>
</thead>
<tbody>
<tr>
<td>AS-CAIDA</td>
<td>16,301</td>
<td>65,910</td>
</tr>
<tr>
<td>Wiki-Vote</td>
<td>7,115</td>
<td>103,689</td>
</tr>
<tr>
<td>CA-GR-QC</td>
<td>5,242</td>
<td>14,496</td>
</tr>
<tr>
<td>CA-HEP-TH</td>
<td>9,877</td>
<td>25,998</td>
</tr>
<tr>
<td>P2P-Gnutella</td>
<td>6,301</td>
<td>20,777</td>
</tr>
</tbody>
</table>

*Source: [http://snap.stanford.edu](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?

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
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