Guidelines for Online Network Crawling: A Study of Data Collection Approaches and Network Properties

Katchaguy Areekijserre
Ricky Laishram Sucheta Soundarajan

Syracuse University, Syracuse, NY, USA

WebSci’18
Data Collection

- The study of complex networks has gained a lot of attention from researchers.
- A convenient way is to get data from APIs.
- Many OSNs provide APIs for accessing data (Facebook, Twitter).

- **Network Sampling / Crawling ≈ Online Sampling**
- **Challenge:** The data collection process takes a lot of time.
- **Question:** Since they are many proposed algorithms, it is often difficult for users to select a crawling technique.
Problem Definition

Let $G = (V, E)$ be a static unobserved, undirected network.

- **Input**: A starting node $n_s$ and query budget $b$.
- In each step, the crawler queries an observed-but-not-queried node. The process repeats for $b$ times.
- **Output**: a sample graph $S = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$, containing all nodes and edges observed.

Two different crawling goals:

1. **Node Coverage**: Maximize a number of observed nodes ($|V'|$).
2. **Edge Coverage**: Maximize a number of observed edges ($|E'|$).

**Related Application**: preserving community structure [MBW10a], preserving high centrality nodes [MBW10b].
Contributions

- Examine how the network properties affect the crawling methods’ performance.
- Perform extensive, scientific analysis of the relationship between network structural properties and the algorithms performance.
- Provide guidelines on how to select an appropriate crawling method.
Observation

**Colored node**: all the sampled nodes.
- Observed-queried nodes
- Observed-unqueried nodes

**Uncolored node**: unobserved nodes.
Hypothesis

- It may be difficult for a crawler to move between regions.
- The crawler gets stuck in one general area. So, it will eventually start seeing the same nodes and edges over and over again (diminishing returns).
Network Properties of Interest

We are interested in 3 properties.

1. Community Separation - Community Mixing/Modularity
2. Node Average Degree
3. Average Community Size

* We select these properties based on the intuition that a crawler has difficulty in moving between regions.
Online Crawling Approaches

We select nine popular algorithms from the literature and categorize them into three classes (G1-G3) based on the results.

- **G1: Node Importance-based Methods**
  - Maximum Observed Degree [ABN+14]
  - Maximum Observed PageRank [SRR12]
  - Online Page Importance Computation [APC03]

- **G2: Random Walk** [LF06]

- **G3: Graph Traversal-based Methods**
  - Breadth-first Search [MMG+07]
  - Depth-first Search
  - Snowball Sampling [AHK+07]
  - Random Crawling
  - Volatile Multi-armed Bandit [BPSF13]
Experiment Studies

We perform two sets of studies.

1. The effects of network properties
   - Controlled experiments on synthetic (LFR model) and real networks.

2. Categorizing network types
   - Studies the algorithms’ performance on different types of networks.
   - collaboration, web, scientific, technological, Facebook, OSNs.
Study 1: The Effects of Network Properties

Results on networks with different values of community mixing $\mu$, average degree $=15$ and average community size $= 300$

Finding

The performance of G1 methods improves as the value of community mixing increases. Others are stable.
Study 1: The Effects of Network Properties

Networks with different values of $d_{avg}$ and $CS_{avg}$ when community mixing $\mu=0.1$.

Finding

- G1 works great on networks with large community sizes.
- G3 performance increases when average degree increases.
- G2 is not affected by these properties.
On real world networks, the performance of methods in G2 drops when modularity increases.
# Study 1: Summary

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Property</th>
<th>G1: Node Importance-Based</th>
<th>G2: Random Walk</th>
<th>G3: Graph Traversal-Based</th>
</tr>
</thead>
<tbody>
<tr>
<td>Node</td>
<td>Community Separation</td>
<td>Excellent performance when community overlap is high (i.e. low $Q$ or high $\mu$).</td>
<td>Stable</td>
<td>Stable</td>
</tr>
<tr>
<td></td>
<td>Average community size</td>
<td>Strong performance when communities are large if $\mu$ is low. Community size does not matter if $\mu$ is high.</td>
<td>Stable</td>
<td>Performance improvement when average degree increases.</td>
</tr>
<tr>
<td></td>
<td>Average degree</td>
<td>Strong performance when average degree is extremely low (&lt;10) even if $\mu$ is low. Otherwise, stable</td>
<td>RW</td>
<td>BFS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Best Method in Group</th>
<th>MOD</th>
<th>RW</th>
<th>BFS</th>
</tr>
</thead>
</table>
**Study 2: Network Types**

The network properties are not known beforehand. How can one select an appropriate method?

<table>
<thead>
<tr>
<th>Type</th>
<th>Network</th>
<th>$d_{avg}$</th>
<th>$CS_{avg}$</th>
<th>$Q$</th>
<th>Properties</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collab.</td>
<td>Citeseer</td>
<td>7.16</td>
<td>988.35</td>
<td>0.90</td>
<td>Low degree, medium-sized and clear communities</td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td>Dblp-2010</td>
<td>6.33</td>
<td>739.91</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dblp-2012</td>
<td>6.62</td>
<td>1248.35</td>
<td>0.82</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>MathSciNet</td>
<td>4.93</td>
<td>594.09</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Recmd.</td>
<td>Amazon</td>
<td>2.74</td>
<td>272.44</td>
<td>0.99</td>
<td>Low degree, small and clear communities</td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td>Github</td>
<td>7.25</td>
<td>83.68</td>
<td>0.43</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FB</td>
<td>OR</td>
<td>25.77</td>
<td>1074.44</td>
<td>0.63</td>
<td>High degree, large and clear communities</td>
<td>G2</td>
</tr>
<tr>
<td></td>
<td>Penn94</td>
<td>65.59</td>
<td>2186.11</td>
<td>0.49</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wosn-friends</td>
<td>25.77</td>
<td>856.65</td>
<td>0.63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tech.</td>
<td>P2P-gnutella</td>
<td>4.73</td>
<td>1276.76</td>
<td>0.50</td>
<td>Low degree, large and clear communities</td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td>RL-caida</td>
<td>6.37</td>
<td>856.12</td>
<td>0.86</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Web.</td>
<td>Arabic-2005</td>
<td>21.36</td>
<td>115.86</td>
<td>1.00</td>
<td>High degree, medium-sized and clear communities</td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td>Italycnr-2000</td>
<td>17.36</td>
<td>1134.34</td>
<td>0.91</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sk-2005</td>
<td>5.51</td>
<td>338.22</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Uk-2005</td>
<td>181.19</td>
<td>157.13</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OSNs.</td>
<td>Slashdot</td>
<td>10.24</td>
<td>173.87</td>
<td>0.36</td>
<td>High degree, small-to-medium-sized and fuzzy communities</td>
<td>G1</td>
</tr>
<tr>
<td></td>
<td>Themarker</td>
<td>29.87</td>
<td>458.90</td>
<td>0.31</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>BlogCatalog</td>
<td>47.15</td>
<td>1455.48</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scientific</td>
<td>PKUSTK13</td>
<td>68.73</td>
<td>3,514.56</td>
<td>0.88</td>
<td>High degree, large and clear communities</td>
<td>G2</td>
</tr>
<tr>
<td></td>
<td>PWTK</td>
<td>51.89</td>
<td>4,635.81</td>
<td>0.93</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shipsec1</td>
<td>24.36</td>
<td>4,117.50</td>
<td>0.89</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shipsec5</td>
<td>24.61</td>
<td>5,252.15</td>
<td>0.90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conclusion

- We performed a large-scale, comprehensive study to understand how the structural features of networks affect the performance of sampling methods.
- Three network properties of interest: community separation, community size, and average degree.
- Algorithm performance is highly dependent on the network structure, and in particular, whether the crawler is able to transition between different regions of the graph.
Thank You

Questions?

kareekij@syr.edu

Yong-Yeol Ahn, Seungyeop Han, Haewoon Kwak, Sue Moon, and Hawoong Jeong, *Analysis of topological characteristics of huge online social networking services*, International conference on WWW, 2007.


References III
