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

Katchaguy Areekijseree
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 $\approx$ 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:

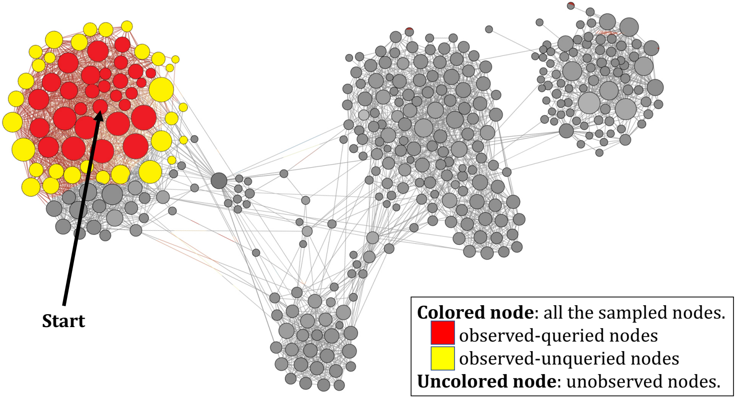
1. Node Coverage: Maximize a number of observed nodes ($|V'|$).
2. Edge Coverage: Maximize a number of observed edges ($|E'|$).

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



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.

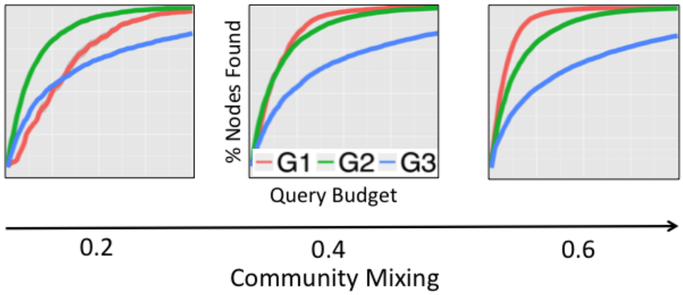
- **G1: Node Importance-based Methods**
  - Maximum Observed Degree [ABN+14]
  - Maximum Observed PageRank [SRR12]
  - Online Page Importance Computation [APC03]
- **G2: Random Walk** [LF06]
- **G3: Graph Traversal-based Methods**
  - Breadth-first Search [MMG+07]
  - Depth-first Search
  - Snowball 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

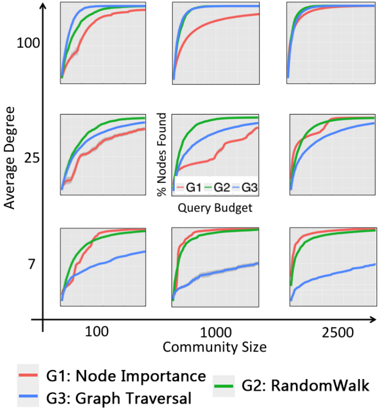


Results on networks with different values of community mixing μ, average degree =15 and average community size = 300

Finding

The performance of G1 methods improves as the value of community mixing increases. Others are stable.

Study 1: The Effects of Network Properties



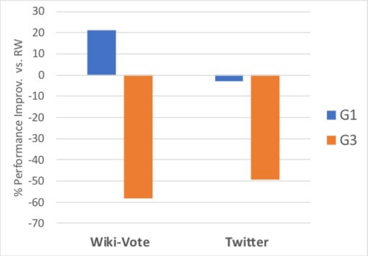
Networks with different values of d_{avg} and CS_{avg} when community mixing μ=0.1.

Finding

- G1 works great on networks with large community sizes.
- G3 performance increases when average degree increases.
- G2 is not affected by these properties.

Study 1: Real World Networks

Test Prop.	Pair	Network	d_{avg}	CS_{svg}	Q
Q	A	Wiki-Vote	28.51	1,177.67	**0.42**
		Twitter	33.01	1,129.25	**0.81**



On real world networks, the performance of methods in G2 drops when modularity increases.

Study 1: Summary

Coverage	Property	G1: Node Importance-Based	G2: Random Walk	G3: Graph Traversal-Based
Node	Community Separation	Excellent performance when community overlap is high (i.e. low Q or high μ).	Stable	Stable
	Average community size	Strong performance when communities are large if μ is low. Community size does not matter if μ is high.		
	Average degree	Strong performance when average degree is extremely low (<10) even if μ is low. Otherwise, stable		Performance improvement when average degree increases.
Best Method in Group		MOD	RW	BFS

Study 2: Network Types

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

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

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

References I

- Konstantin Avrachenkov, Prithwish Basu, Giovanni Neglia, Bruno Ribeiro, and Don Towsley, *Pay few, influence most: Online myopic network covering*, Computer Communications Workshops, 2014.
- 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.
- Serge Abiteboul, Mihai Preda, and Gregory Cobena, *Adaptive on-line page importance computation*, Proceedings of the 12th international conference on World Wide Web, 2003.
- Zahy Bnaya, Rami Puzis, Roni Stern, and Ariel Felner, *Bandit algorithms for social network queries*, 2013 International Conference on Social Computing, IEEE, 2013, pp. 148–153.

References II

Jure Leskovec and Christos Faloutsos, *Sampling from large graphs*, Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2006, pp. 631–636.

Arun S Maiya and Tanya Y Berger-Wolf, *Online sampling of high centrality individuals in social networks*, Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2010, pp. 91–98.

________, *Sampling community structure*, Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 701–710.

References III

Alan Mislove, Massimiliano Marcon, Krishna P Gummadi, Peter Druschel, and Bobby Bhattacharjee, *Measurement and analysis of online social networks*, 7th ACM SIGCOMM conference on Internet measurement, ACM, 2007, pp. 29–42.

Mostafa Salehi, Hamid R Rabiee, and Arezo Rajabi, *Sampling from complex networks with high community structures*, Chaos **22** (2012), no. 2.