Sampling in networks

Argimiro Arratia & Ramon Ferrer-i-Cancho

Universitat Politècnica de Catalunya

Version 0.6
Complex and Social Networks (2020-2021)
Master in Innovation and Research in Informatics (MIRI)

Instructors

- Ramon Ferrer-i-Cancho, rferrericancho@cs.upc.edu, http://www.cs.upc.edu/~rferrericancho/
- Argimiro Arratia, argimiro@cs.upc.edu, http://www.cs.upc.edu/~argimiro/

Please go to http://www.cs.upc.edu/~csn for all course's material, schedule, lab work, etc.

The "problem" of analyzing networks

Sampling comes to our rescue

A few possible scenarios:

1. We have collected a *large* graph that fits into memory, but want to run an expensive algorithm that may take too long. How can we speed up the computation?
2. We have collected a *huge* graph that fits into disk but not main memory. How can we analyze it in reasonable time?
3. It is extremely costly or impossible to collect the entire graph (think Facebook, WWW, Twitter, etc.), we only have access to subgraphs via *crawling*, and yet we want to infer properties of the underlying graph.

In all of these scenarios, sampling (implicitly or explicitly) is used!

Understanding sampling is important!

A little story of not so long ago..

- 1999-2000: several acclaimed reports on power-law degree distribution of various networks
  - Internet: [Faloutsos et al., 1999]
  - WWW: [Albert et al., 1999]
  - Metabolic networks: [Jeong et al., 2000]
- 2003: it is shown empirically that the *sampling procedure* may induce a power-law, even if the underlying graph is not scale-free! [Lakhina et al., 2003]
- 2005: further empirical and theoretical studies support this [Achlioptas et al., 2005, Clauset and Moore, 2005]

Conclusion: it is very important to understand how biases in sampling affect results

In today's lecture

Sampling strategies

Biases of sampling strategies

Sampling General Goals

How do we measure the goodness of a sample, as well as the method of sampling?

Depends on what do we compare against:

Scale-down goal: We want the sample graph S to have similar properties as the original G

Back-in-time goal: We want the sample graph S to be similar to what G looked like back in the time when it had same size of S

Goals

1. Sample a representative subgraph (scale-down goal)
   - that is, obtain a subgraph that has similar properties, for a set of representative properties *simultaneously* (e.g.: degree distribution, clustering coefficient, community structure, etc.)
2. Estimation of a network parameter (back-in-time goal)
   - E.g.: average degree of nodes, diameter, ...
3. Estimate node attributes (back-in-time goal)
   - E.g.: age of users in a social network
4. Estimate edge attributes (back-in-time goal)
   - E.g.: relationship type of friends in a social network

Different sampling strategies will work for certain goals better than others

Overview of sampling strategies

From [Leskovec and Faloutsos, 2006, Maiya and Berger-Wolf, 2011, Ahmed et al., 2014]

- Random node selection
  - Only possible when access to entire graph is given
- Random edge selection
  - Only possible when access to entire graph is given
- Crawling-based
  - Snowball sampling: BFS, DFS, Forest Fire, ...
  - Random walks

[A spoiler note: For scale-down sampling goal best performers are based on random walks, since these are biased towards high degree nodes and guarantee connectivity. For back-in-time goal: Forest-fire, PageRank sampling of nodes; these mimic the temporal evolution of the graph]

Random node selection

Several possibilities

- Uniform node sampling
- Degree-based sampling [Adamic et al., 2001]
  - Probability of visiting node proportional to its degree (assumed known)
  - Originally used for searching [Adamic et al., 2001]
- Pagerank-based sampling [Leskovec and Faloutsos, 2006]
  - Probability of visiting node proportional to its pagerank (assumed known)

Random edge selection

Several possibilities

- Uniform edge sampling
  - sample edges and then include incident nodes
- Random node-edge sampling
  - select node uniformly at random, then select incident edge uniformly at random
- Hybrid sampling [Krishnamurthy et al., 2005]
  - With probability 0.8, perform random node-edge sampling
  - With probability 0.2, perform uniform edge sampling
- Induced edge sampling [Ahmed et al., 2014]
  - Uniformly sample edges
  - Complete graph sample with edges between nodes incident on sampled edges

Crawling I

a.k.a. "sampling by exploration"

- Breadth-First search (BFS)
  - explore neighbors of least recently visited nodes
- Depth-First search (DFS)
  - explore neighbors of most recently visited nodes
- Random walk (RW) [Gjoka et al., 2010]
  - explore neighbors of most recently visited nodes uniformly at random (no queue)
- Forest Fire sampling (FFS) [Leskovec et al., 2005]
  - probabilistic version of BFS
  - with probability p (typically 0.7), visit neighbor

Crawling II

a.k.a. "sampling by exploration"

- Expansion sampling (XS) [Maiya and Berger-Wolf, 2010, Maiya and Berger-Wolf, 2011]
  - greedily add node maximizing *expansion* $\frac{|N(S)|}{|S|}$
- Random walk with jump (RJ) [Ribeiro and Towsley, 2010]
  - same as random walk, but jump to random node with probaility p

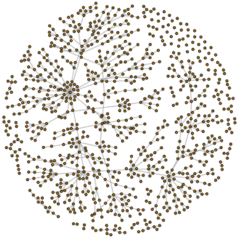
In today's lecture

Sampling strategies

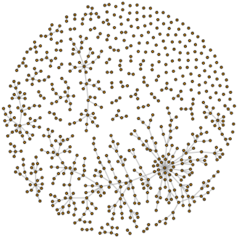
Biases of sampling strategies

Uniform node sampling

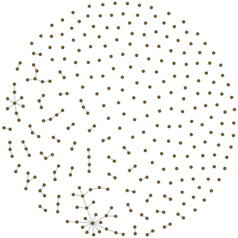
- Induced subgraphs of scale-free networks are not scale-free [Stumpf et al., 2005]
- Induced subgraphs of connected scale-free networks are sparse



90% of nodes



70% of nodes



30% of nodes

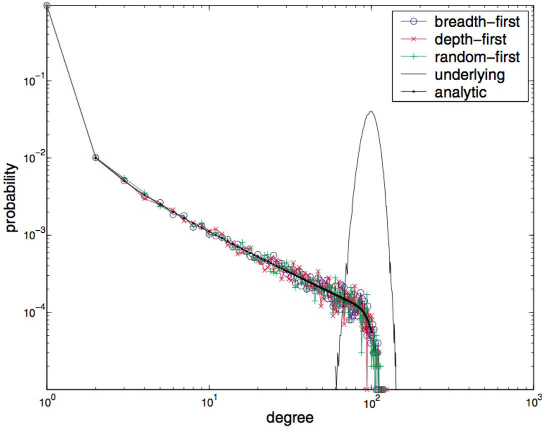
Crawled subsets of ER graphs are scale-free

[Lakhina et al., 2003][Clauset and Moore, 2005]

[Lakhina et al., 2003] observe this empirically by sampling ER-graphs with trace-route routine (a minimum spanning tree)

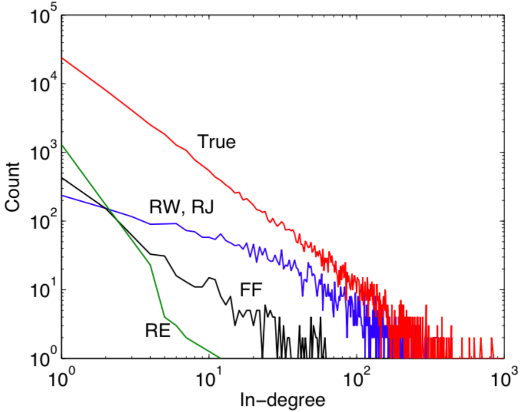
[Clauset and Moore, 2005] Give a general proof of this fact (worth reading!). Basic argument is that traceroutes from single source can be modelled as a spanning tree. Then show that building a spanning tree in Erdos-Renyi graph gives subgraph with degree distribution following a power law of the form $P(k) \approx k^{-1}$

Crawled subsets of ER graphs are scale-free
[Clauset and Moore, 2005]



More crawling biases

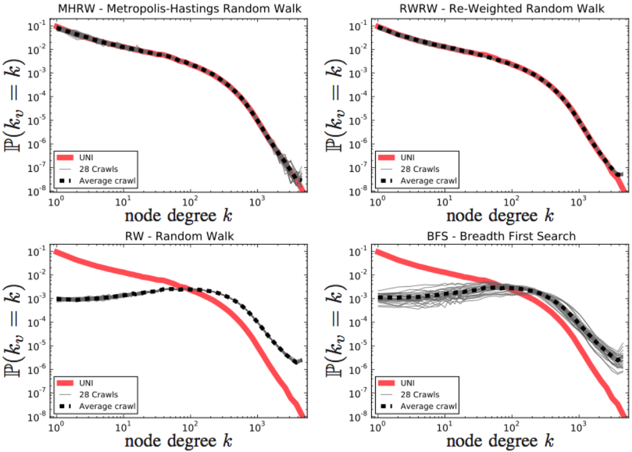
In general, random walks, DFS, and BFS lead to over-sampling of high-degree nodes



Compensating for RW bias

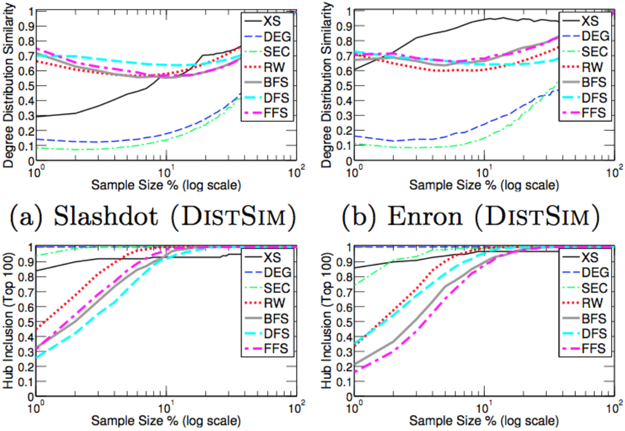
- Random Walk (RW)
  - Nodes with high degree are over-represented since probability of visiting a node $v \propto k_v$
- Re-Weighted random walk (RWRW)
  - Hansen-Hurwitz estimator for non-uniform selection probabilities
  - After the walk, re-weight $\hat{p}(k) = \frac{\sum_{v:k_v=k} 1/k_v}{\sum_v 1/k_v}$
- Metropolis-Hastings random walk (MHRW)
  - Walk with new transition probabilities $P_{v \to w} = \frac{1}{k_v} min(1, \frac{k_v}{k_w})$
  - i.e. select random neighbor, and move with probability $min(1, \frac{k_v}{k_w})$
  - i.e. always accept moves to nodes of lower degree, reject some moves to nodes of higher degree
  - results in uniform probabilities of visiting nodes

Uniform sampling of Facebook users using random walks
[Gjoka et al., 2010]



Results from [Maiya and Berger-Wolf, 2011]

Degree distribution

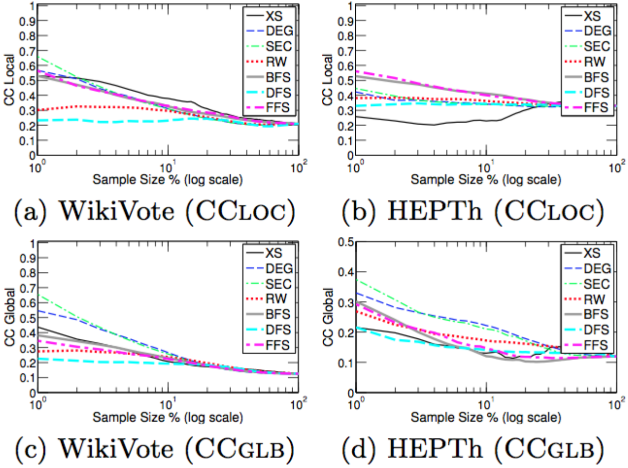


(a) Slashdot (DistSim) (b) Enron (DistSim)

(c) Slashdot (Hubs) (d) Enron (Hubs)

Results from [Maiya and Berger-Wolf, 2011]

Clustering coefficient

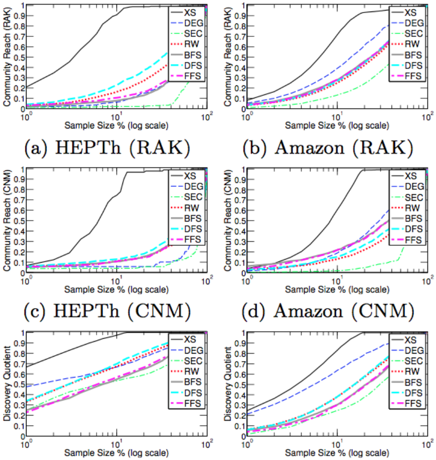


(a) WikiVote (CCLoc) (b) HEPTh (CCLoc)

(c) WikiVote (CCGLB) (d) HEPTh (CCGLB)

Results from [Maiya and Berger-Wolf, 2011]

Network reach



(a) HEPTh (RAK) (b) Amazon (RAK)

(c) HEPTh (CNM) (d) Amazon (CNM)

(e) HEPTh (DQ) (f) Amazon (DQ)

References I

Achlioptas, D., Clauset, A., Kempe, D., and Moore, C. (2005).
On the bias of traceroute sampling: or, power-law degree distributions in regular graphs.
In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 694–703. ACM.

Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. (2001).
Search in power-law networks.
Phys. Rev. E, 64:046135.

Ahmed, N., Neville, J., and Kompella, R. (2014).
Network sampling: From static to streaming graphs.
ACM Trans. Knowl. Discov. Data, to appear.

References II

Albert, R., Jeong, H., and Barabási, A.-L. (1999).
Internet: Diameter of the world-wide web.
Nature, 401(6749):130–131.

Clauset, A. and Moore, C. (2005).
Accuracy and scaling phenomena in internet mapping.
Physical Review Letters, 94(1):018701.

Faloutsos, M., Faloutsos, P., and Faloutsos, C. (1999).
On power-law relationships of the internet topology.
SIGCOMM Comput. Commun. Rev., 29(4):251–262.

References III

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010).
Walking in facebook: A case study of unbiased sampling of osns.
In *INFOCOM, 2010 Proceedings IEEE*, pages 1–9. IEEE.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabasi, A. L. (2000).
The large-scale organization of metabolic networks.
Nature, 407(6804):651–654.

References IV

Krishnamurthy, V., Faloutsos, M., Chrobak, M., Lao, L., Cui, J. H., and Percus, A. G. (2005).
Reducing large internet topologies for faster simulations.
In *Proceedings of the 4th IFIP-TC6 International Conference on Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communication Systems*, NETWORKING'05, pages 328–341, Berlin, Heidelberg. Springer-Verlag.

References V

Lakhina, A., Byers, J. W., Crovella, M., and Xie, P. (2003).
Sampling biases in IP topology measurements.
In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 1, pages 332–341.

Leskovec, J. and Faloutsos, C. (2006).
Sampling from large graphs.
In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631–636. ACM.

References VI

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005).
Graphs over time: Densification laws, shrinking diameters and possible explanations.
In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, KDD '05, pages 177–187, New York, NY, USA. ACM.

Maiya, A. S. and Berger-Wolf, T. Y. (2010).
Sampling community structure.
In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 701–710, New York, NY, USA. ACM.

References VII

Maiya, A. S. and Berger-Wolf, T. Y. (2011).
Benefits of bias: Towards better characterization of network sampling.
In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 105–113. ACM.

Ribeiro, B. and Towsley, D. (2010).
Estimating and sampling graphs with multidimensional random walks.
In *Proceedings of the 10th ACM SIGCOMM Conference on Internet Measurement*, IMC '10, pages 390–403, New York, NY, USA. ACM.

References VIII

Stumpf, M. P. H., Wiuf, C., and May, R. M. (2005).
Subnets of scale-free networks are not scale-free: Sampling properties of networks.
Proceedings of the National Academy of Sciences of the United States of America, 102(12):4221–4224.