Predicted Max Degree Sampling: Sampling in Directed Networks to Maximize Node Coverage

Ricky Laishram, Katchaguy Areekijseree, Sucheta Soundarajan
Department of EECS, Syracuse University, NY
{rlaishra, kareekij, susounda}@syr.edu

Motivation

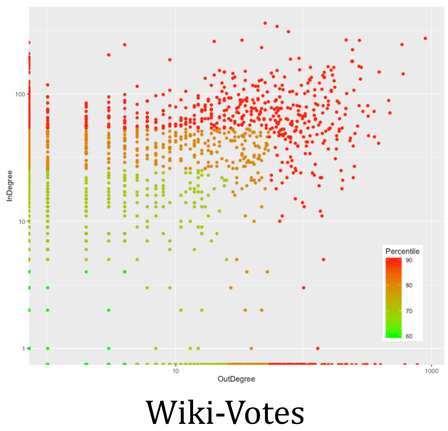
- Exploration in many real world networks are possible only thorough crawling.
- There are usually costs associated with each query during the crawling process.
- **If we have a limited amount of query, we need a efficient algorithm to find nodes in the network.**
- In undirected networks, the strategy of querying the node with the maximum observed degree works well.
- However, **In directed networks, there is very little correlation between the in-degree and out-degree of the high degree nodes.**
- So, we need a better algorithm to efficiently find nodes in the directed networks with the given budget.

Problem Definition

- **Given:**
  - A directed network $G = \langle V, E \rangle$ that can only be explored through crawling.
  - An initial sub-graph $G_0^* = \langle V_0^*, E_0^* \rangle$.
  - A query budget B.
- **Goal:**
  - **Obtain a sub-graph that maximizes the number of observed nodes.**
- **Assumptions:**
  - We can query for either the in-neighbors (Γ_i) or/and out-neighbors (Γ_o) of an observed node.
  - Each query cost one unit of the budget.

Degree Correlation

Web-Stanford

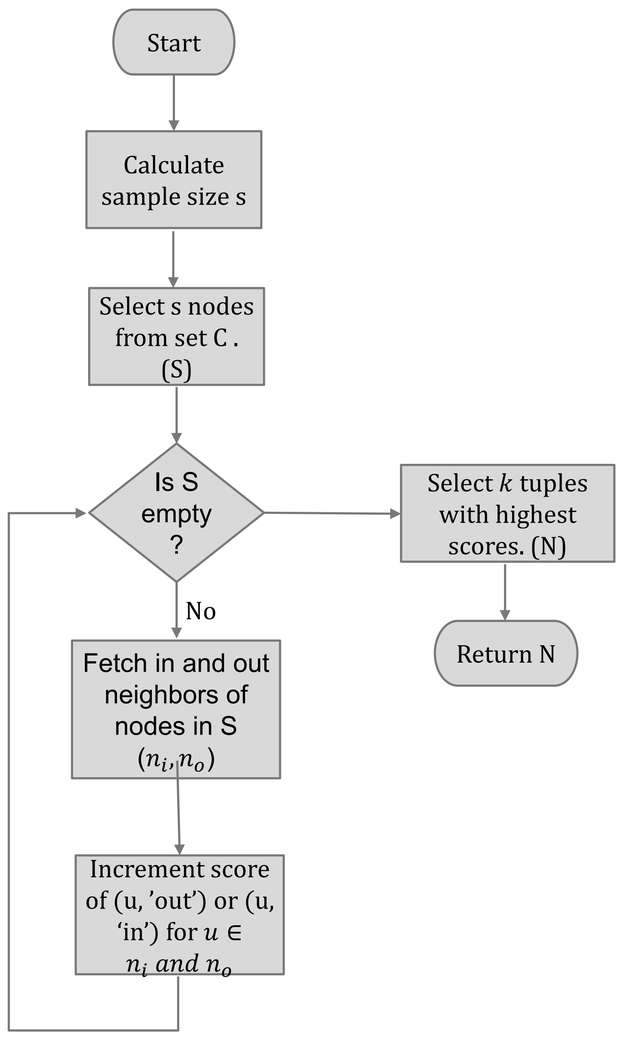


Wiki-Votes

Intuition

- **In-Closed Nodes (C^i)** and **Out-Closed Nodes (C^o)** are the set of nodes for which we know all the in-neighbors and out-neighbors respectively. $C = C^i \cup C^o$
- **Open Nodes (O)** are the set of nodes that have been observed but not in C.
- Query a sample of the closed nodes, and find k open nodes (and query type) which was observed most frequently. Call this set N.
- Perform the appropriate query on the nodes in N.

Predicted Max Degree



- Sample size:

$$\prod_{i=1}^{d_\phi}(|C| + 1 - s - i) \leq (1-p)\prod_{i=1}^{d_\phi}(|C| + 1 - i)$$

$$s \in \mathbb{Z}^+$$

- Accuracy:
  - $a = \frac{|\{(u,\tau) \in N : degree_\tau(u) \geq d\_\phi\}|}{|N|}$
- If $a \geq p$, increase k for the next iteration. Otherwise decrease k.

R Laishram, K Areekijseree, S Soundarajan. "Predicted Max Degree Sampling: Sampling in Directed Networks to Maximize node coverage", IEEE BigData 2016.

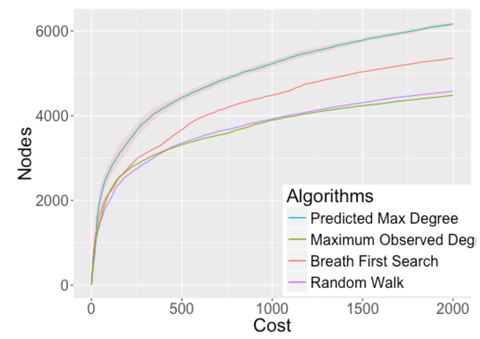
Experimental Setup

- Datasets:

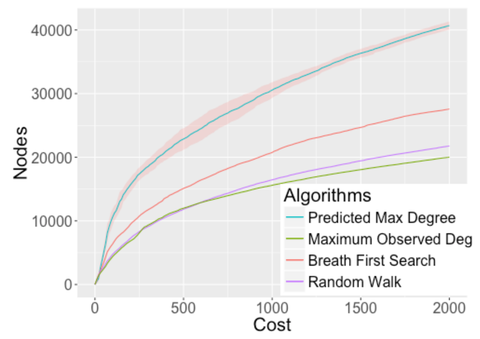
	Nodes	Edges
Wiki-Votes	7115	103689
Soc-Slashdot	82168	948464
Web-Google	875713	5105309

- $p = 0.9, \phi = 90$
- Baseline:
  - BFS
  - Random Walk
  - MOD

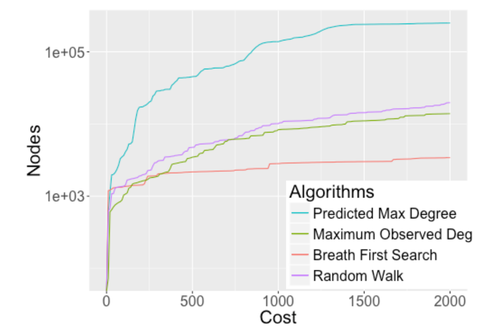
Experimental Results



- Wiki-Votes



- Soc-Slashdot



- Web-Google

Conclusion

- We propose a new algorithm to sample a directed network with the goal of maximizing the node coverage within a given budget.
- **We show experimentally that our algorithm performs better than the baseline algorithms for all the datasets we considered.**