Lab #11: Graph Sampling

Everything Data

CompSci 216 Spring 2015

DUKE
COMPUTER SCIENCE

Announcements (Mon. Apr. 6)

- Feedback on your **project mid-term report** to be emailed by this weekend
- Winners of Lab #9 (visualization)
  - **Most thought-provoking & novel**: Team #1

    Arihant Jain, Jennie Ju, Arun Karottu, Devin Solanki
  - **Best design**: Team #13

    Jackson Borchardt, David Clancy, Anthony Hagouel, David Maydew, Ezgi Ustundag

A quick review

Useful stats when getting to know a graph:

- # of nodes and # of edges
- Degree distribution
- Distance distribution
- Clustering coefficient distribution
- PageRank distribution
- Betweenness distribution

Sampling

- Use a fraction of the available data to make inferences about the whole dataset

Why?

- Dataset is too large to deal with
  - Acquiring all data is too expensive
  - Analyzing all data is too slow

Graph sampling

- Given an input graph $G = (V, E)$, construct a subgraph $H = (V', E')$, where $V' \subseteq V$ and $E' \subseteq E$

- Want H to have the "same properties" as G

Edge sampling

- Choose *E'* as a random sample of *E*
  - A node is picked if an incident edge is picked
- Resulting graph can be very sparse
  - Large diameter, small clustering coefficient
- Nodes with higher degrees in *G* have a higher chance of being picked
  - Not a random sample on nodes
  - Degree distribution is not preserved

Node sampling

- Choose V' as a random sample of V
  - An edge is picked if its endpoints are picked
- Degree distribution looks similar
- Resulting graph can be sparse
  - Clustering coefficient is smaller, because when nodes are sampled uniformly at random, they are less likely to form triangles

Random jump sampling

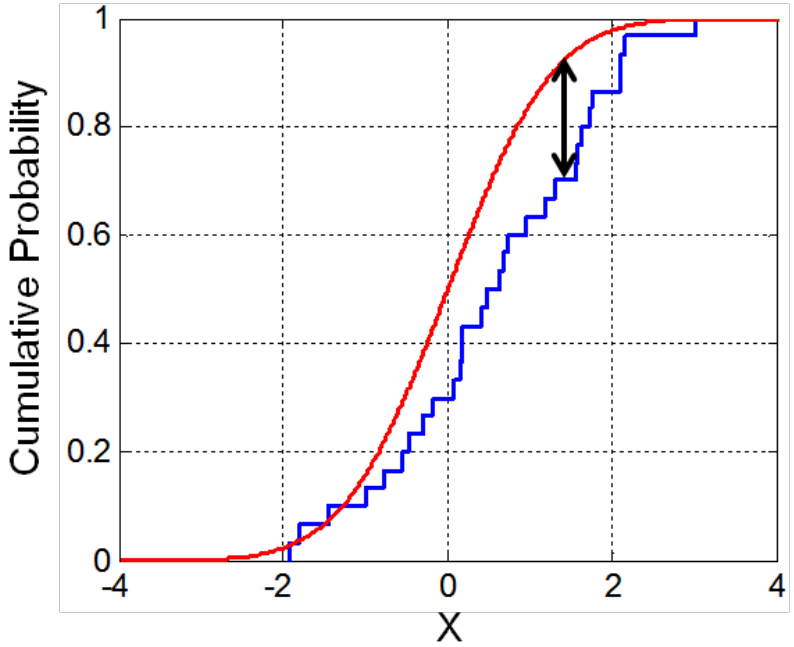
- Pick a starting node uniformly at random
- Perform a random walk with teleporting
  - With probability d, choose an outgoing edge uniformly at random
  - With probability $(1 - d)$, teleport to a random starting node and restart the random walk
- Stop when enough number of nodes are visited; compute the induced subgraph
  - An edge is picked if its endpoints are picked

Random jump sampling

- If $d = 0$, then it is node sampling!
- With $d > 0$, biased towards high-degree nodes
  - Not a random sample on nodes
  - Degree distribution is not preserved
- Preserves the clustering coefficient distribution
  - Because random walk visits connected components

Measuring sample quality

- D-statistic (aka Kolmogorov-Smirnov statistic) that computes the difference between two distributions
  - Smaller means closer
  - 0 means identical



http://en.wikipedia.org/wiki/File:KS_Example.png

Lab challenge

- Implement your sampling procedure to downsize a 4000-node Facebook graph to a 400-node one
- Preserve the distributions of degrees and clustering coefficients as much as possible
  - Random jump sampling is a good start; feel free to use new ideas
  - Beat average D-statistic of 0.25 for extra credit
  - Lowest average D-statistic wins the challenge