Complex-Network Modelling and Inference

Lecture 23: Network Sampling

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
https://roughan.info/notes/Network_Modelling/

School of Mathematical Sciences,
University of Adelaide

March 7, 2024

Section 1

Network Sampling

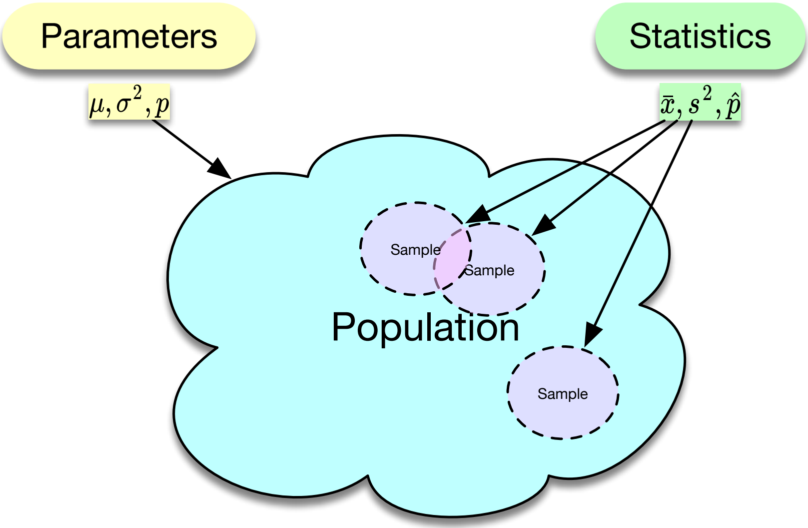
Why sample

- Some graphs are very big!
  - measurements cost (money, time, resources, ...)
  - maybe too big to analyse
- Some measurement approaches can't help it
  - missing data is common
  - missing data creates a kind of sampling
- Visualisation

Sampling goals

The goal of sampling is to obtain a reasonably accurate measure of the particular statistics of the overall population.

- Your definition of "reasonable" may vary
- The statistics you are interested in will vary
  - statistics of the nodes, or edges, or triangles, ...
    - remember, they represent people, or relationships, ...
  - network metrics (we spent 3 lectures on these)
  - model parameters (we spent even more time on models)



(figure stolen from Jono)

Notes

- We could be
  - sampling some graphs from a larger set
  - *sampling some part of a single graph*
- Properties of interest
  - *unbiased*: expected value of estimator is the same as the statistic, *e.g.*, $\mathbb{E}[s] = \sigma$
  - *asymptotically unbiased:* the above is true as the number of samples increases (convergence in expectation)
  - *consistent*: estimates converge in probability
  - *efficient*: MSE of estimate is as small as possible for the number of samples
- Assume uniquely labelled nodes
  - so we can tell if we hit the same node twice
  - sometimes say a node is "burned" if already sampled
  - can have a method that "re-samples" nodes deliberately (not my most favoured idea though)

Problems

- Bias in general
  - if we preferentially sample some subgroup we can easily introduce bias into our statistics estimate
  - ideally, we would have random samples to avoid this
- Structural bias
  - in our problems, the population members are not independent, they have relationships
  - so we don't just need random sample of the population, we also need (somehow) to see a random view of their relationships
- Some properties are properties of the whole graph
  - Hamiltonian and Eulerian cycles
  - k-connectivity
- We presume that we must sample without knowledge of the underlying graph
  - if you know the graph, why sample?

Sampling strategies

Somewhat mirror measurement strategies

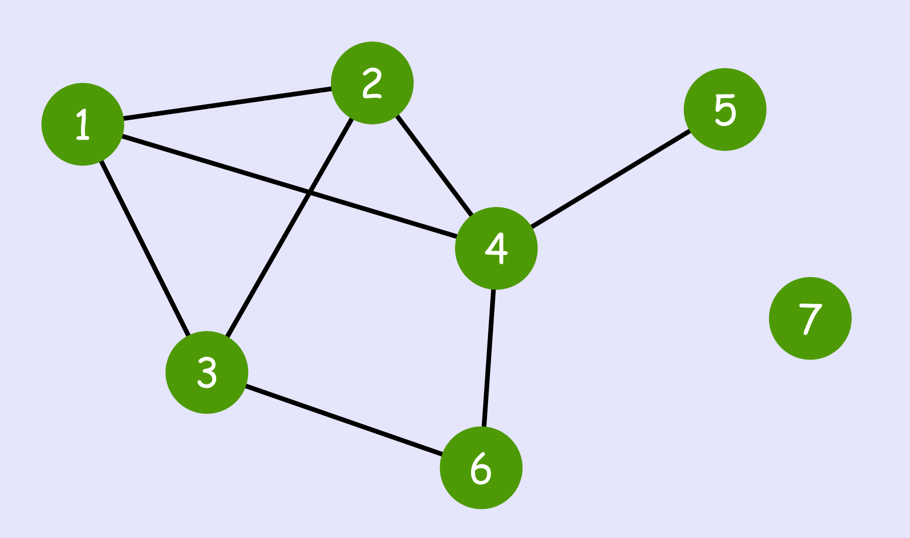
- Node sampling
- Edge sampling
- Random-walk sampling
- Snowball sampling
- Path-based sampling

Node sampling

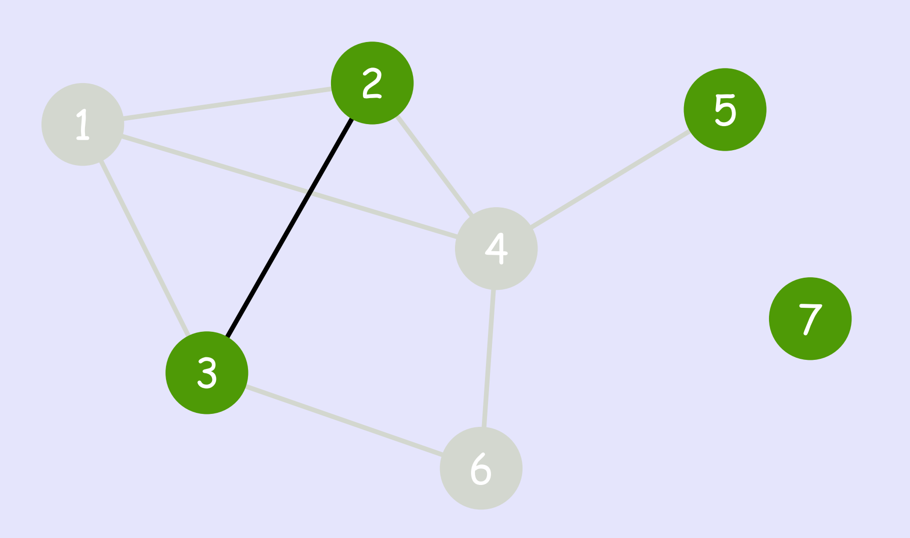
Graph $G(N, E)$

- Randomly choose a subset of nodes $N' \subset N$
  - *e.g.*, randomly generate a Facebook ID, and see if it is real
- Choose $E' \subset E$, such that all edges between nodes in N' are in E'

Node Sampling Example



Node Sampling Example



Node Sampling Pros and Cons

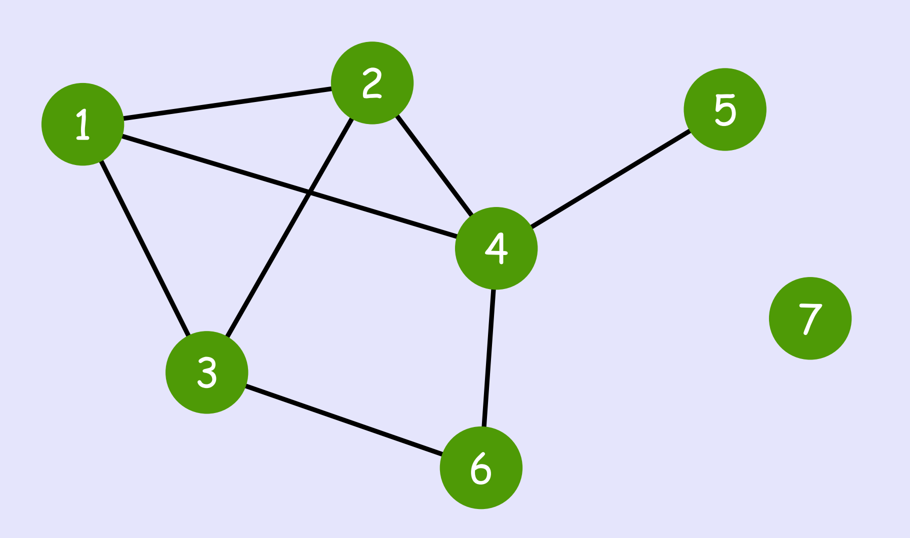
- Pros:
  - simple
  - unbiased sample of nodes
    - sampled GER random graph will be a GER random graph
- Cons:
  - sparsifies the network
    - Q: is the node degree you measure the degree in the subgraph, or the degree of the sampled nodes in the original graph?
  - breaks the structure, *e.g.*,
    - clustering coefficient will be smaller
    - breaks up connected components
    - distances will be longer
  - not easy to get an unbiased sample of nodes in many situations

Edge Sampling

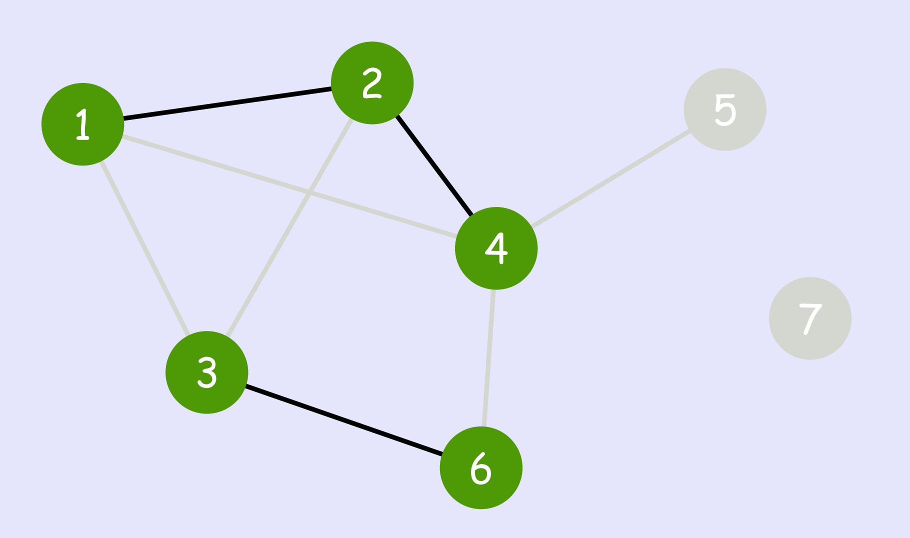
Graph $G(N, E)$

- Randomly choose a subset of edges $E' \subset E$
- Choose $N' \subset N$, such that all end-points of edges in E' are in N'

Edge Sampling Example



Edge Sampling Example



Edge Sampling Pros and Cons

- Pros:
  - simple
  - unbiased sample of edges
  - properties such as assortativity preserved
- Cons:
  - biased sample of nodes, *e.g.*,
    - preferentially samples nodes with high degree
    - don't see nodes with zero degree
  - also breaks structure of network
  - not all networks can be measured/sampled this way

Weighting

- With either of the above we could weight the sample
    - sample as before
    - accept/reject with probability dependent on node/edge features
    - *e.g.*, sampling with weight depending on centrality of node
    - not obvious how to do it without introducing biases, without knowing something about the network *a priori*

Random-walk sampling (with escaping)

- Pick a random start
- Perform a random walk from each seed
  - probability d keep going
  - probability $1 - d$ pick a new random start point
- Stop when "enough" nodes are sampled

Alternative is Frontier Sampling [RT10] – start from a set of random seeds, and process the RWs in parallel

Random-walk sampling Pros and Cons

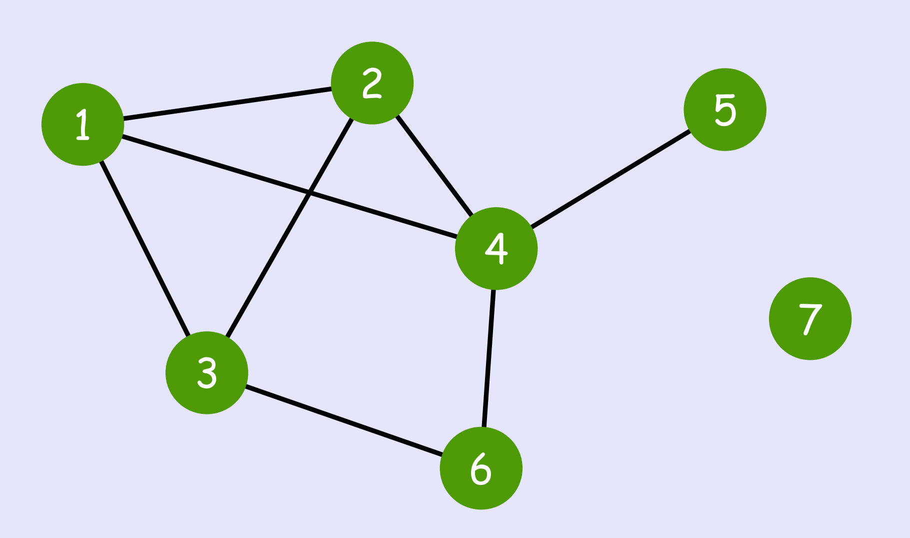
- Pros:
  - uniform distribution on edges
  - preserves clustering (better than other approaches), and some other properties
- Cons:
  - biased towards higher degree nodes

Snowball Sampling [Col58]

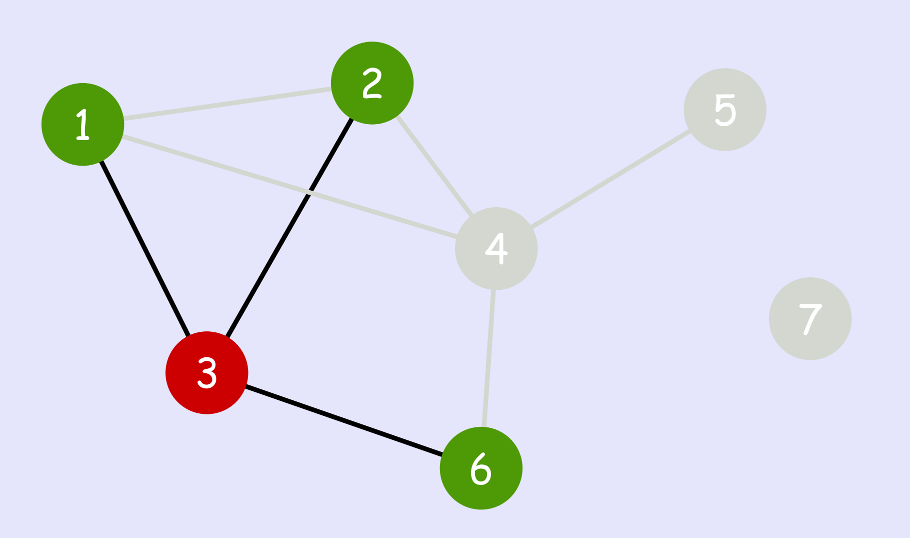
- Sample some seed nodes
- Include their neighbours, and their neighbour's neighbours out to some number of hops
  - might be a sub-sample of neighbours
  - might be a fixed number of neighbours
  - links might be suggested by survey respondent

Variants are called "chain-referral" or "network" or "forest-fire" sampling.

Snowball Sampling Example



Snowball Sampling Example



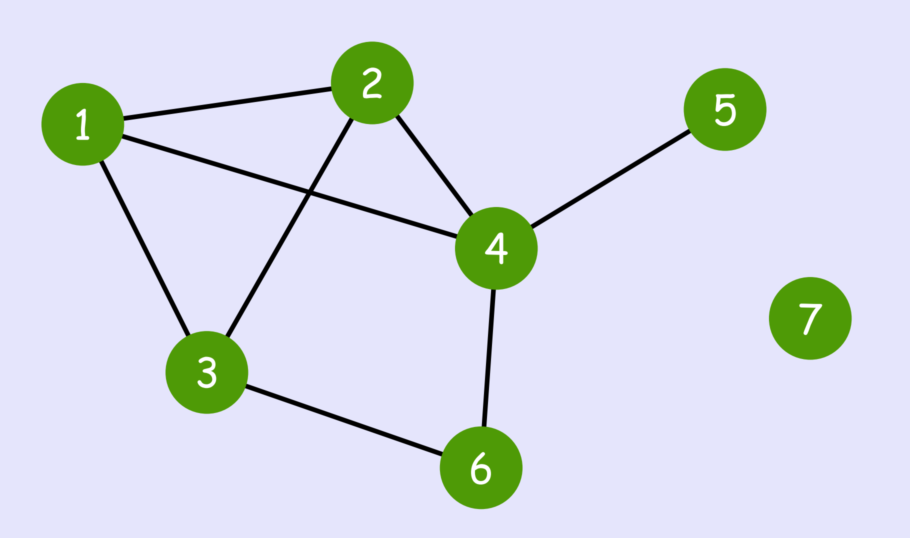
Snowball Sampling Pros and Cons

- Pros:
  - often driven by practicalities of measurements
    - it can be hard to “find” a set of original nodes to sample
  - preserves local structure
- Cons:
  - inefficient if sampling rate is high (get overlaps)
  - biased selection of nodes (and edges)
  - only preserves local structure
  - can make network look MORE clustered

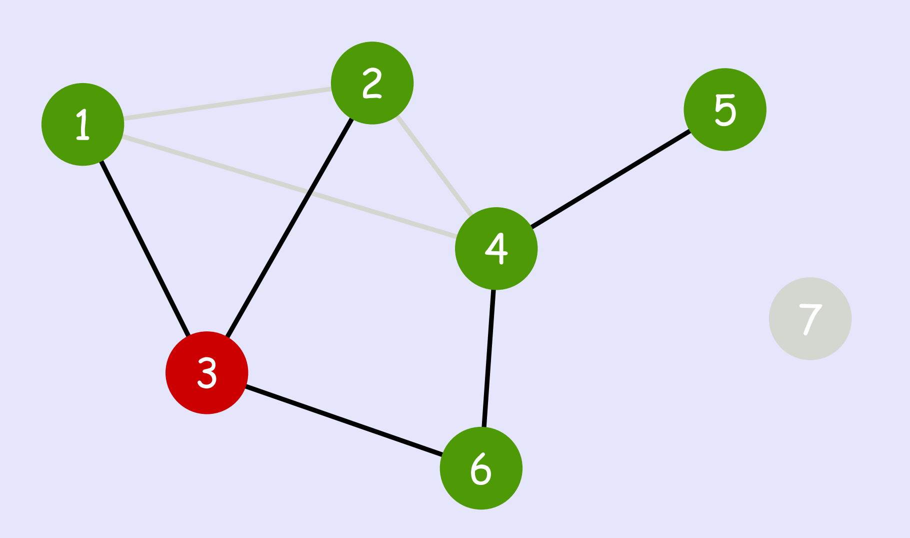
Path-based Sampling

- Start from a (hopefully) random seed
- Follow the shortest path tree away from the node
  - follow the used pathways

Path-based Sampling Example



Path-based Sampling Example



Path-based Sampling Pros and Cons

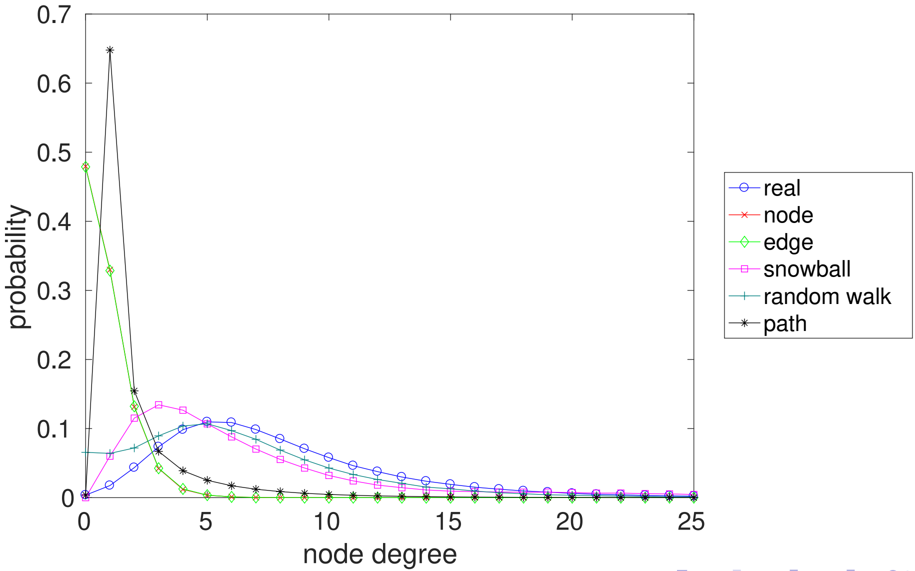
- Pros:
  - often driven by practicalities of measurements
  - preserves distances
- Cons:
  - inefficient if sampling rate is high (get overlaps)
  - introduces unexpected biases, *e.g.*, degree distribution, that can be extreme [LBCX03, ACKM09]

The degree of distortion depends on the model

- GER random graph
  - 10,000 nodes
  - $\bar{k} = 8$
- generate and sample 100 instances
- sampling rates
  - node: $1/10$ nodes
  - edge: $1/10$ edges
  - snowball: 2 seeds, 3 hops
  - random walk: $d = 0.15$, $1/10$ nodes
  - path: 1 seed, all (connected) destinations

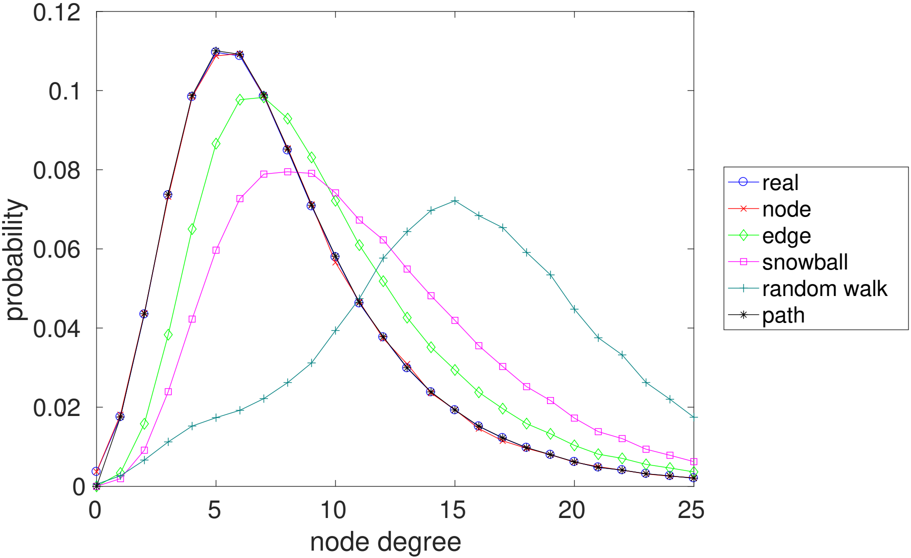
Degree distributions

Degree of nodes in the sampled subgraph



Degree distributions (2)

Degree of sampled nodes in the original graph



Clustering

sample method	global clustering
node	0.0042
edge	0.0003
snowball	0.0118
random walk	0.0265
path	0.0000
unsampled	0.0038

Yet More Sampling Strategies

- Path-, Random-Walk and Snowball are all traversal sampling strategies, there are others
  - Metropolis-Hastings Random Walk
- ???

A Few More Bits

- There is no perfect solution here – all methods introduce some type of bias, or break something
- Given a model, and a sampling strategy, we can sometimes reverse sampling biases
  - derive distributions analytically
  - invert
  - but not guaranteed to be possible as there is some information loss
- Haven't really considered difference for directed graphs

Further reading I

Dimitris Achlioptas, Aaron Clauset, David Kempe, and Cristopher Moore, *On the bias of traceroute sampling: Or, power-law degree distributions in regular graphs*, J. ACM **56** (2009), no. 4, 21:1–21:28.

James Coleman, *Relational analysis: The study of social organizations with survey methods*, Human Organization **17** (1958), no. 4, 28–36.

Pili Hu and Wing Cheong Lau, *A survey and taxonomy of graph sampling*, CoRR **abs/1308.5865** (2013).

Anukool Lakhina, John Byers, Mark Crovella, and Peng Xie, *Sampling biases in IP topology measurements*, IEEE Infocom, April 2003.

Sang Hoon Lee, Pan-Jun Kim, and Hawoong Jeong, *Statistical properties of sampled networks*, Phys. Rev. E **73** (2006), 016102.

Further reading II

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