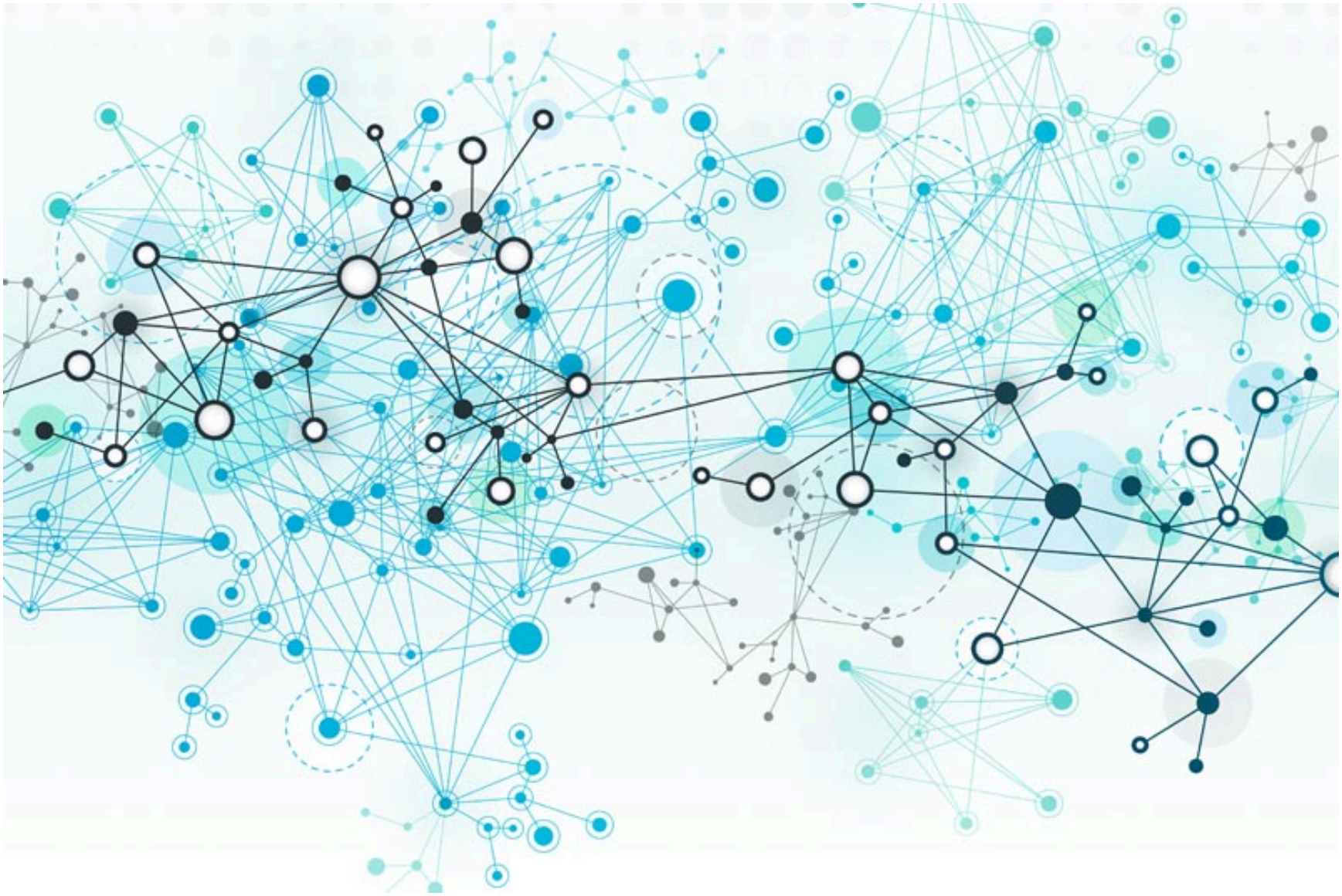
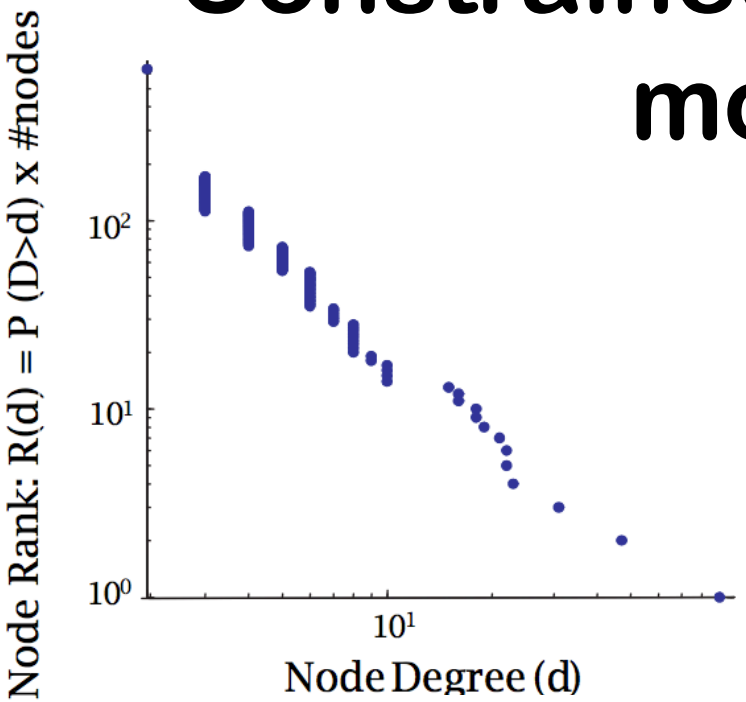
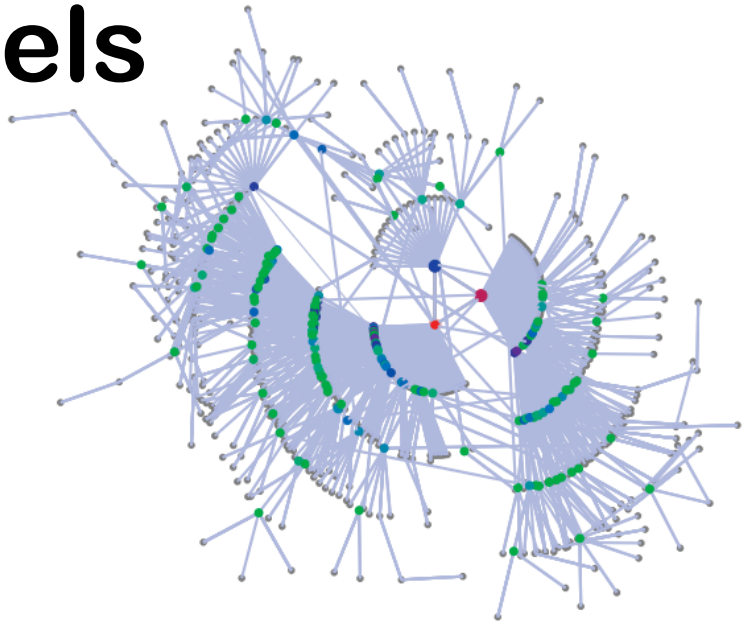
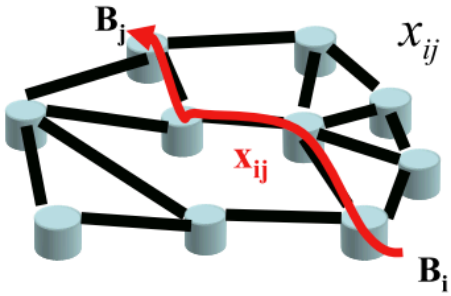
BIG NETWORK DATA
ECE 289 UC San Diego



Constrained optimization models





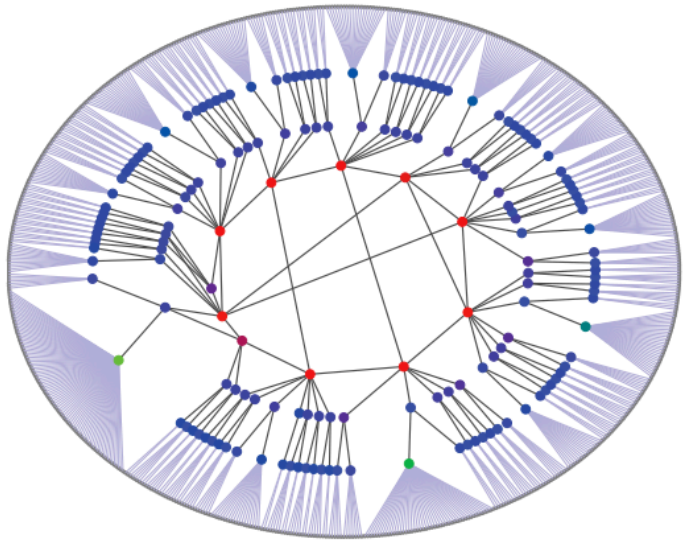


$$x_{ij} \propto B_i B_j$$

$$\max_{\alpha} \sum_{i,j} x_{ij} = \max \sum_{i,j} \alpha B_i B_j$$

$$s.t. \sum_{i,j:k \in r_{ij}} x_{ij} \le B_k, \forall k$$

(a)



(b)

Figure 5. Generating networks using constrained optimization. (a) Engineers view netw

Constrained optimization models

Start with a low-degree low-diameter “backbone”

Attach “tree-like” regional points of access

Determine the routing matrix through shortest-path algorithm

Solve constrained flow optimization problem

The value for the obtained flows is higher than the one obtained solving the same problem using a PA generative model

The reason is because the network model reflects real engineering insights

Constrained optimization models

- The backbone cannot be composed of highly connected hubs
- Router capacity is limited by bandwidth and number of incident links
- We cannot build fast switches serving a large number of ports
- Traffic aggregation is performed at the edges of the network where high degree nodes (ISP) provide low bandwidth connections to end users
- Whether degree distribution follows a power law or not is inconsequential

Power law degree distribution

3Faloustos (1999) observed that the Internet graph both at the router level (IP) and at the AS level (BGP) has a power law degree distribution

$$P(k_i = k) = Ak^{-\alpha}, \quad 2 < \alpha < 3$$

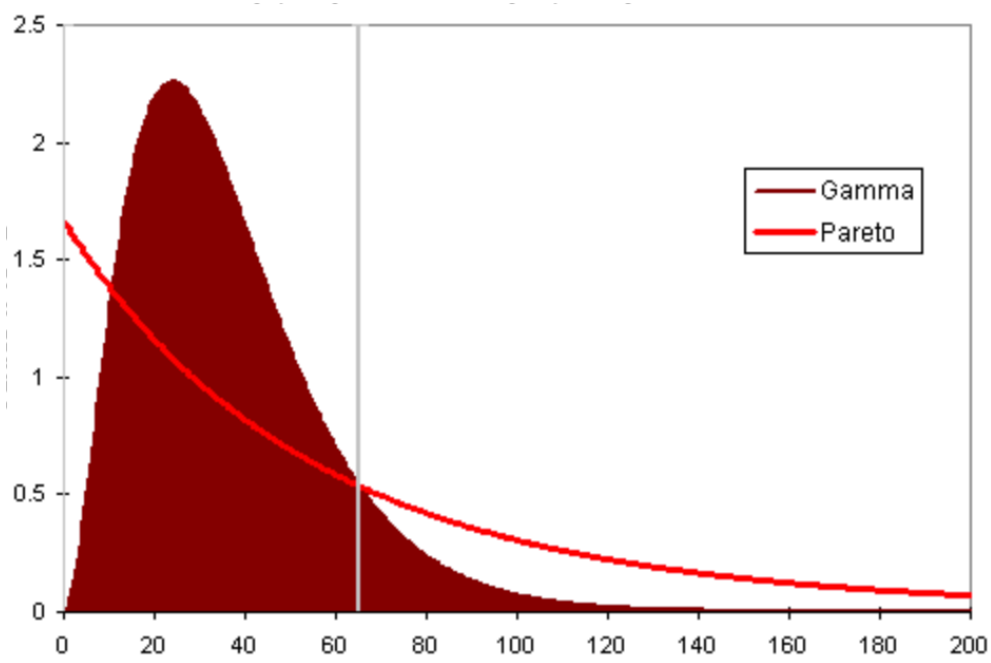
This indicates large variability in node degrees, as the average node degree is essentially uninformative

This distribution is incompatible with classic random graph (ER) models

Power law degree distribution

This distribution is compatible with preferential attachment (PA) models, but these models fail to capture other properties of the real system

Models based on constrained optimization might be more relevant



Heuristically optimized trade-offs

Consider a random tree driven by a uniform distribution of points in the unit square

$$i \leftrightarrow j : \min_{j<i} \quad \alpha d_{i,j} + h_j$$

Every newly added node minimizes the weighted sum of two objectives

"Last Mile" connection cost (Euclidean distance)

"Centrality" (Hop-distance to other nodes)

$h_j = \mathbb{E}(\text{hops to others})$

$h_j = \max(\text{hops to others})$

$h_j = \text{hops to central node}$

Heuristically optimized trade-offs

Fabrikant, Koutsoupias, Papadimitriou (2002)

$$\alpha < 1/\sqrt{2} \implies \text{T is a star}$$

$$\alpha > c_1\sqrt{n} \implies \mathbb{E}(|\{i : \deg_i \geq k\}|) < n^2 \exp(-c_2 k)$$

$$\alpha > 4,\ \alpha = o(\sqrt{n}) \implies \mathbb{E}(|\{i : \deg_i \geq k\}|) > c(k/n)^{-\alpha}$$

Heuristically optimized trade-offs

This suggests that power laws can be the manifestation of trade-offs, complicated optimization problems with multiple and conflicting objectives.

Finding the correct trade-offs requires an understanding of these complex processes that drive the network construction mechanism

Multi-objective optimization

Mandelbrot (1953)

Design of the “optimal” language

Determine the set of frequencies assigned to n words maximizing information transmission and minimizing transmission cost

$$f_1 \leq f_2 \leq \ldots \leq f_n$$

i-th word of length $\log i$

$$\max \frac{-\sum f_i \log f_i}{\sum f_i \log i}$$

Optimal frequencies follow a power law!

Power law degree distribution

3Faloustos (1999) observed that the Internet graph both at the router level (IP) and at the AS level (BGP) has a power law degree distribution

Barabasi et al. quickly followed proposing PA as a “universal” model for complex networks exhibiting power laws but providing unreliable predictions

Many other models are possible and more appropriate that reflect the complex trade-offs required in engineering design not captured in a rich-gets-richer process

But is it really a power law?

Unreliable measurements

IP Alias resolution problem

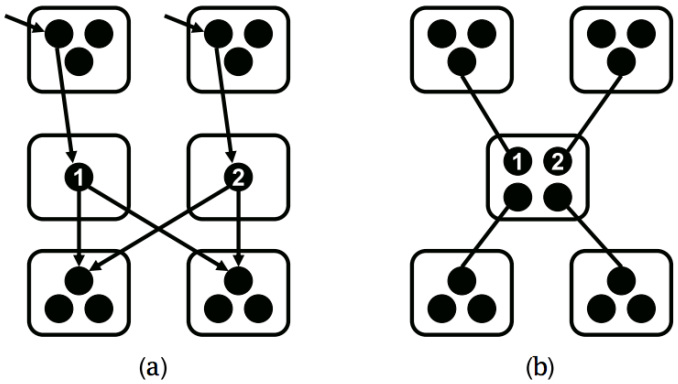
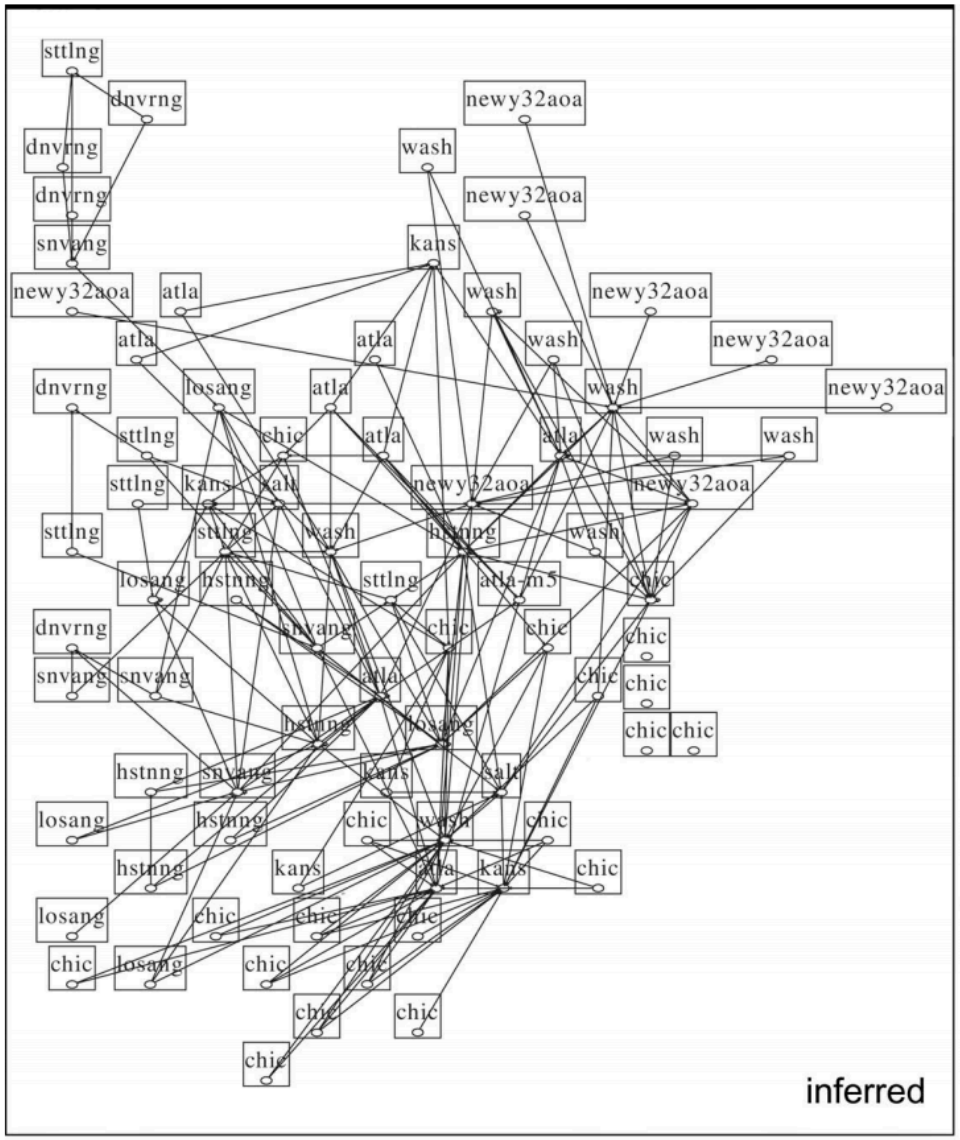
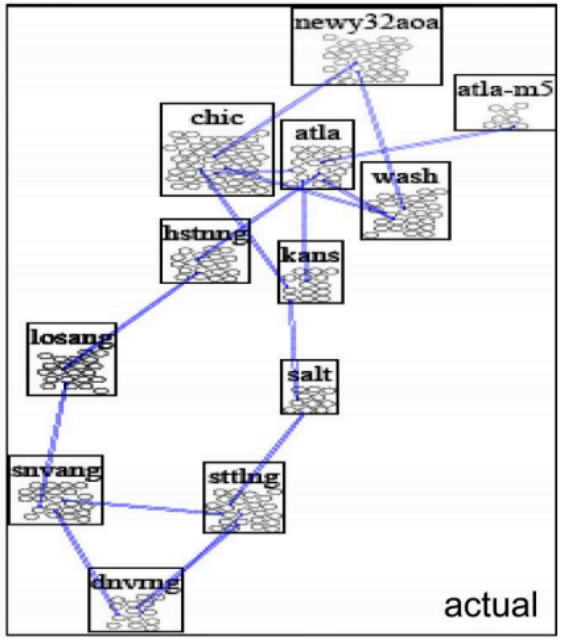
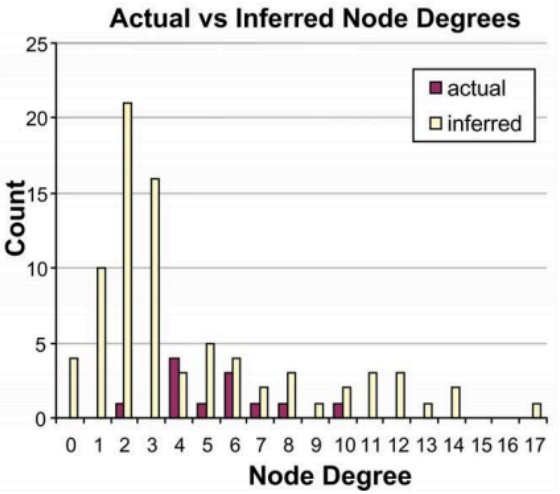


Figure 2. The IP alias resolution problem.

Unreliable measurements

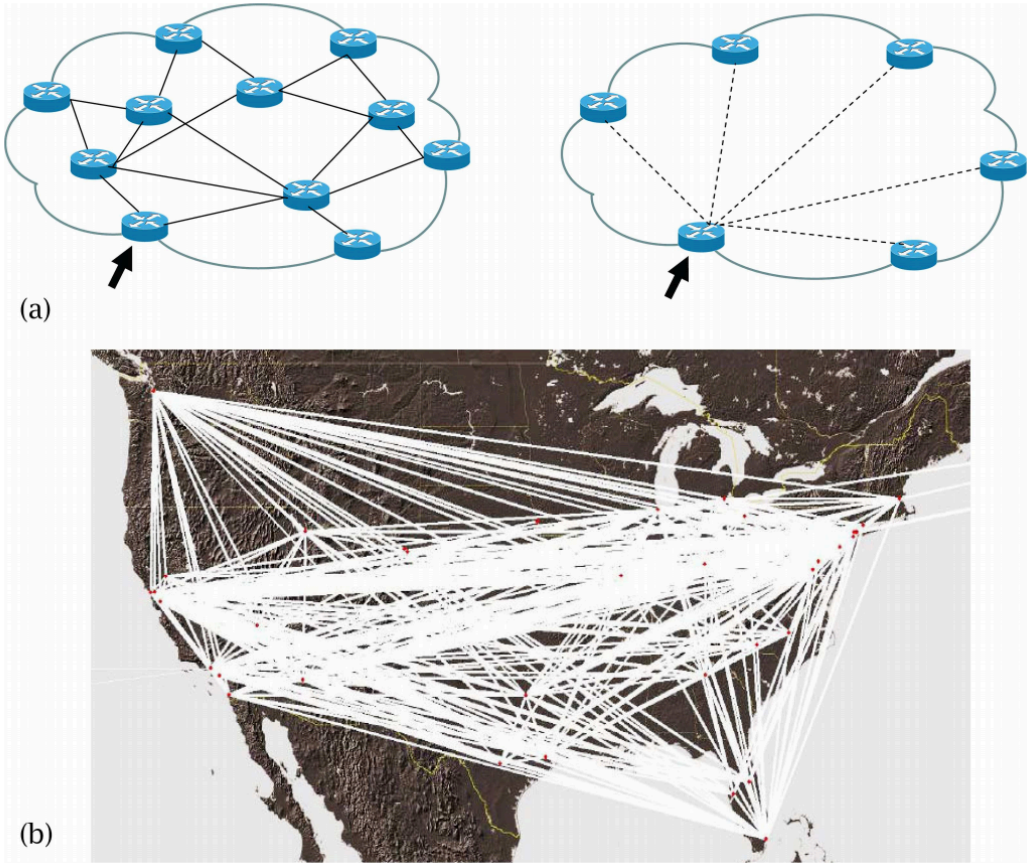






Unreliable measurements

Hidden layer network problem



Biased sampling

Even if we assume measurements are reliable and we sample a BSF tree

High degree nodes and nodes close to the root are more likely to be sampled

Sampling is biased with respect to the property to be sampled!

The bias introduced by BSF sampling can make power laws appear where they do not exist. Even a random (ER) graph or a regular random graph, where each vertex has the same degree is reported to have a power law degree distribution.

Lakhina et al. (2003)
Clauset and Moore (2005)
Achlioptas et al. (2005)

Biased sampling

How can we infer the true degree distribution from sampling?

This is a key problem in network science, beyond internet modeling, for example in social networks exploration. Methods usually involve some amount of bias that needs to be controlled.

Biased sampling

Random walk methods

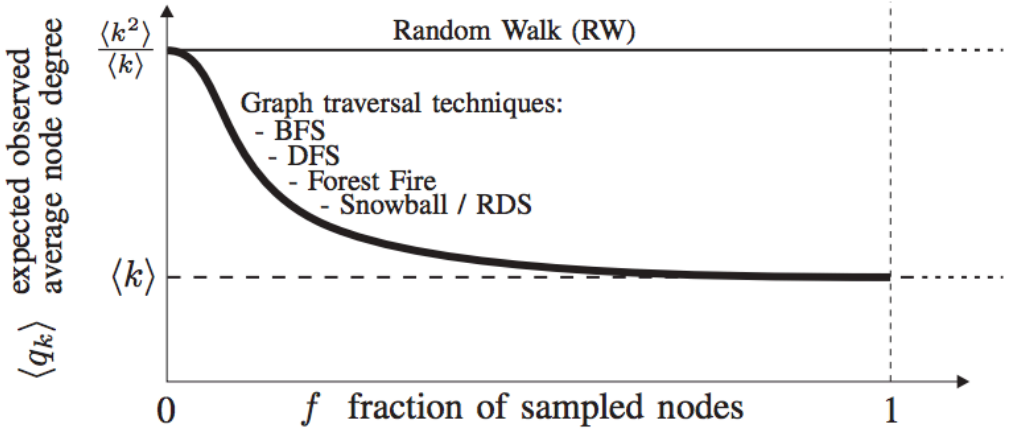
- nodes can be traversed multiple times

Graph Traversal methods

- Each node is visited at most once
- Tree methods, all nodes sampled to construct tree, some edges are seen (e.g. traceroute method described before)
- Sub-graph methods, some nodes are sampled, all incident edges are seen

Biased sampling

Given a random graph of given degree distribution, what is the observed distribution as a function of the fraction of sampled nodes?



Kurant et al. (2011)

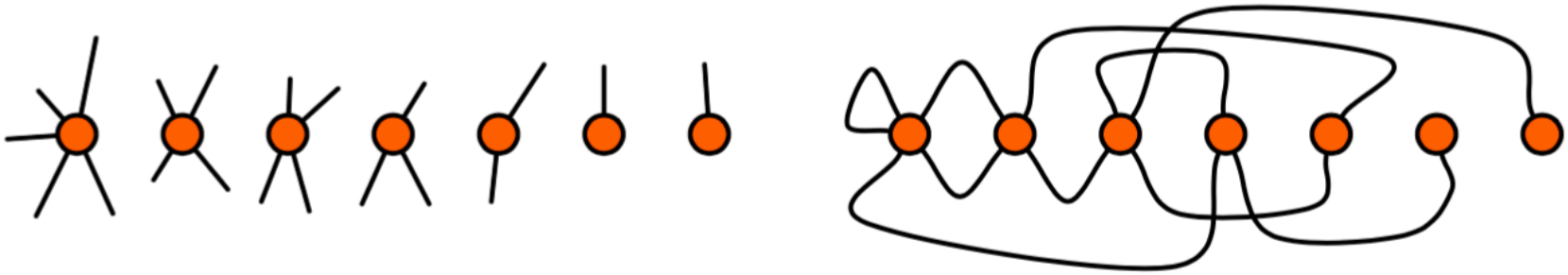
Fig. 1. **Overview of results.** In this paper, we calculate the node degree distribution q_k expected to be observed by BFS in a random graph $RG(p_k)$ with a given degree distribution p_k, as a function of the fraction of sampled nodes f. In this plot, we show only the average $\langle q_k \rangle$. We show the RW as a reference. $\langle k \rangle = \langle p_k \rangle$ is the real average node degree, and $\langle k^2 \rangle$ is the real average squared node degree. *Observations:* (i) For a small sample size, BFS has the same bias as RW; with increasing f, the bias decreases; a complete BFS ($f = 1$) is unbiased. (ii) All graph traversal techniques (that use sampling without replacement) lead to the same bias in $RG(p_k)$. (iii) The shape of the BFS curve depends on the graph (the real node degree distribution p_k), but it is always monotonically decreasing; we calculate it precisely in this paper. (iv) We also correct for the bias and compute the original distribution p_k based on the sampled q_k and f (not shown here).

For the configuration model

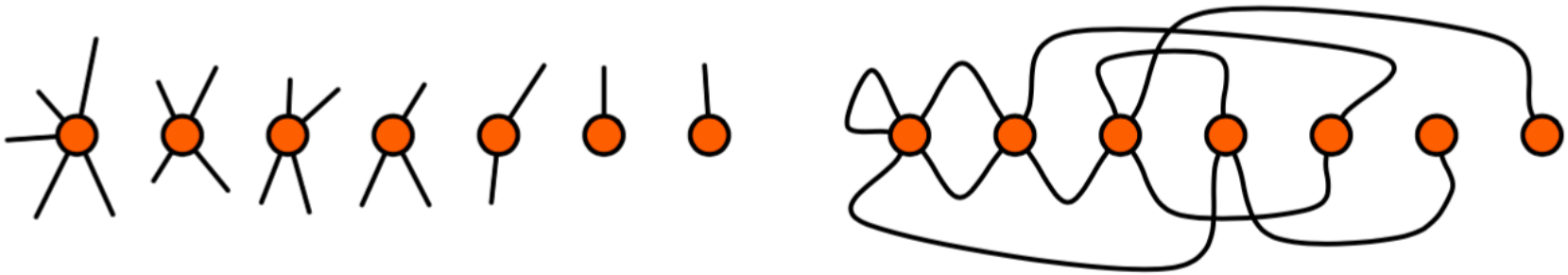
Every node is given a number of “stubs” according to a given degree distribution (fraction of nodes with degree k)

$$\sum_{v \in V} k_v = 2|E|$$

Stubs are randomly matched in pairs until |E| edges are created



For the configuration model



Expected observed degree distribution after exploring a fraction f of nodes

$$q_k(f) = \frac{p_k(1-(1-t(f))^k)}{\sum_l p_l(1-(1-t(f)^l)}$$

See details in the paper available on-line