Complex networks
Sampling

János Török

Department of Theoretical Physics

May 15, 2023

Sampling Networks

- Why?: Performance, and time limitation
- Reason:
  - Actual limit in the resources
  - Test ideas fast
  - Limited access
  - Temporal access
- How?: Depends what you want, but always complicated

Based on the lecture of Mohammad Al Hasan, Nesreen K. Ahmed, Jennifer Neville, Purdue University, West Lafayette, IN

Network characteristics

- Task: Measure should give the same value on the sampled network than on original:
- Measure type:
  - Single node: e.g. degree distribution, average degree
  - Link correlations: e.g. centrality, assortativity, clustering
  - Mesoscopic correlations: e.g. community structure, motifs
- Different level of correlations require different approaches
- Single node properties are the easiest to retain

Sampling scenarios

- Full access to the network
- Restricted access (through a collection of seed nodes)
- Streaming access (data not sampled is lost forever) (Not covered here)

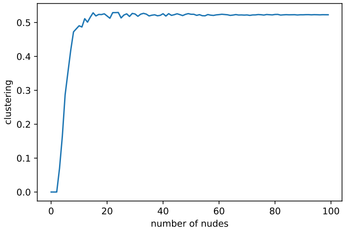
Full access, only nodal attributes

- Uniform node sampling
- Degree base random node sampling
- Random pagerank sampling
- Random edge sampling

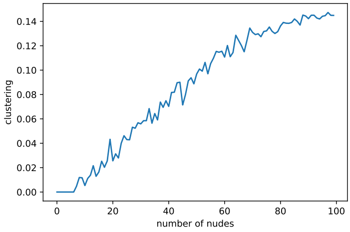
Random node sampling

- Uniform node selection
- Conserved quantities
  - Average degree
  - Average of any nodal attribute
  - Any function of nodal attributes (e.g. degree distribution)
- Quantities not conserved
  - Multi nodal correlations are systematically destroyed

SBM



BA



Degree based random node sampling

- Node selection is proportional to function $\pi(k)$ of node degree
- Bias to nodes with higher degree
- Use case
  - Degree distribution is generally decreasing
  - Few large degree nodes are generally not selected by random node selection, for which measures have high error for large degrees
  - If degree distribution and $\pi(k)$ is known sampled estimates can be corrected.
- Generally $\pi(k) = k$

Degree based random node sampling

- Very often conditional averages are calculated and contition is on degree, (e.g. assortativity)
- Select few nodes with each representative degree
- Problems:
  - High error for low degree nodes (e.g. error goes as $\sim 1/\sqrt{k}$): oversample low degree nodes accordingly (rule of thumb same amount of cpu time for each bin)
  - Sproadic k values for large degree: allow range for large degree nodes anyway the error in degree will still be small
  - Feel free to drop irrelevant degrees (e.g. for humans $50 < k < 500$)

Pagerank based random node sampling

- Node selection is proportional to Pagerank probability $dk_{in}/M + (1-d)/N$
- The previous two can be obtained as a special case with $d = 0$ and $d = 1$
  - Small degree nodes have tunable probability to be selected
  - Measured quantities can be transferred back to original system

Random edge sampling

- Uniform edge selection
- A vertex is selected in function of the degree of the vertex u

$$P = 1 - (1 - \rho)^{k(u)}$$

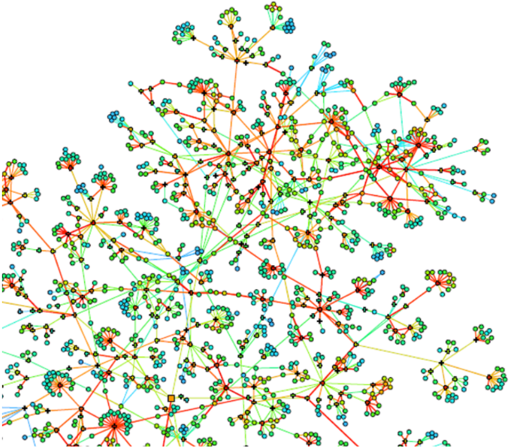
- For $\rho \to 0$, $P(k) = \rho k$
- For $\rho \ll 0$ bias is reduced
- Edge statistics are conserved
- Nodal statistics will be biased to high-degree vertices

Sampling under restricted access

- There are few (or 1) entry points
- No global property is known a priori
- Network supports crawling, neighbors of accessed nodes are known
- Graph traversal methods
  - Snowball sampling
  - Breadth-First Search
  - Depth-First Search
  - Forest fire
- Random walk based methods
  - Classic random walk
  - Random walk with restart
  - Markov Chain Monte Carlo using Metropolis-Hastings algorithm

Snowball sampling

- Start from a seed
- Sample all links to neighbors
- (In some version this step is limited to *n* neighbors)
- Visit all neighbors and there also sample all links to neighbors
- Stop at desired level



Snowball sampling

- Start from a seed
- Sample all links to neighbors
- Visit all neighbors and there also sample all links to neighbors
- Stop at desired level
- Advantage: simple, and long history in social science
- Problems:
  - Non random
  - Last layer has almost always degree 1
  - For large degree only very few layers can be sampled, very often two

Snowball sampling: Variations

- Breadth-first Sampling:
  - Above version
  - Discover vertices at distance d before discovering any at distance $d + 1$
- Depth-first Sampling:
  - Discover farthest vertex along a chain
  - If there is no more than go back recursively
- Forest Fire Sampling
  - Neighbors of the current node are added with probability p
  - The above is repeated until some condition
  - Note the forest fire may go extinct before it reaches the desired number of nodes or depth
- $n-$Snowball sampling
  - For the each active node discover only n neighbors
  - A node can be chosen if it has not been visited before

Random walk

- Start from a seed
- Do a random walk
- All links to the visited node are discovered
- Biased towards high degrees
- Samples the current community much more than the rest of the network (can be a desired effect)

Random walk with restart

- Start from a seed
- Do a random walk
- All links to the visited node are discovered
- Biased towards high degrees
- With probability d jumps back to origin
- Samples the current community much more than the rest of the network, even more than simple random walk
- Could be useful if one wants a good sample of a community from an otherwise enormous network

Markov Chain Monte Carlo using Metropolis-Hastings algorithm

- Correct the random walk bias
- Go to a node with probability depending on the degree of the target node
- Current node i, target node j

$$P(i \to j) = \min(k_i/k_j, 1)$$

- Thus we always go towards smaller degree nodes but only with probability k_i/k_j towards larger degree ones
- In theory this model gives uniform sampling of the nodes

Horovitz-Thompson estimator

- Calculate the mean μ of a quantity X_i over the finite set S of nodes.
- If sampling is unbiased of course we have

$$\mu = \frac{1}{|S|} \sum_{i \in S} X_i,$$

  where $|S|$ is the cardiality of the set S
- If there is a bias π_i for selecting node i (of course π can also be a function of X and other quantities)
- The Horovitz-Thompson estimator:

$$\mu_{HT} = \frac{1}{|S|} \sum_{i \in S} X_i / \pi_i$$

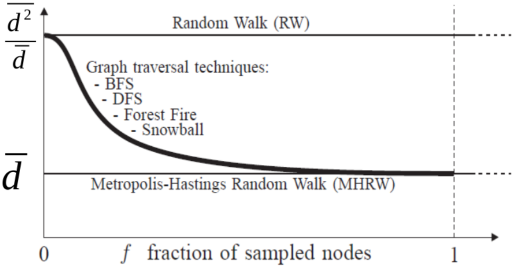
Vertex selection probability (bias)

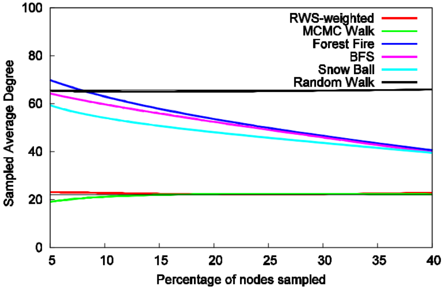
- Note: in image $d \equiv k$ the degree of a node

Method	Vertex Selection Probability, $\pi(u)$ $\lvert V\rvert = n, \lvert E\rvert = m,$
RN, MH-uniform target	$\frac{1}{n}$
RDN, RWS	$\frac{d(u)}{2m}$
RPN, RWJ	$c \cdot \frac{d_{in}(u)}{m} + (1-c) \cdot \frac{1}{n}$ (undirected) $c \cdot \frac{d(u)}{2m} + (1-c) \cdot \frac{1}{n}$ (directed)
RE	$\sim \frac{d(u)}{2m}$
RNE	$\frac{1}{n}\left(1 + \sum_{x \in adj(u)} \frac{1}{adj(x)}\right)$

Vertex selection probability (bias)

- Note: in image $d \equiv k$ the degree of a node





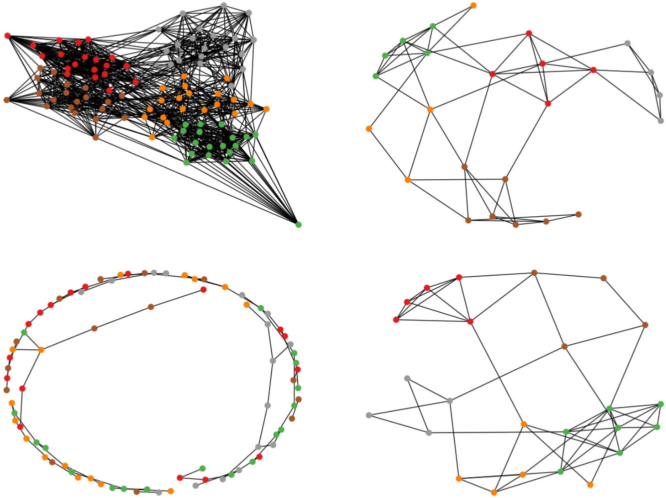
Full access neighbor correlations

- Using all methods the clustering coefficient will be wrong
- This is because the triangles are missing, and have low probability
- Solution: Induction
  - Include links between sampled nodes
- Partial induction
  - Include links between sampled nodes With probability p
- Note: nomenclature
  - *induced:* all links between selected nodes (e.g. egocentric network)
  - *incident:* all edges between nodes of selected links

Samples: 25% of the nodes

original

random node

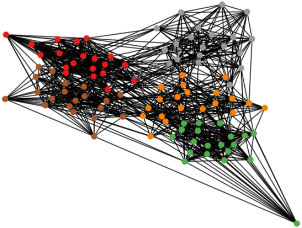


random edge

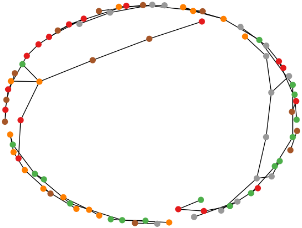
pagerank

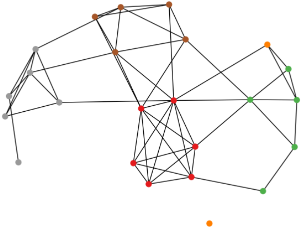
Samples: 25% of the nodes

original

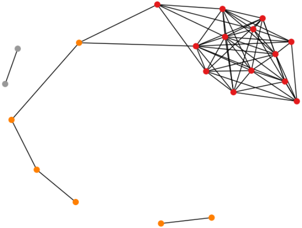


random edge





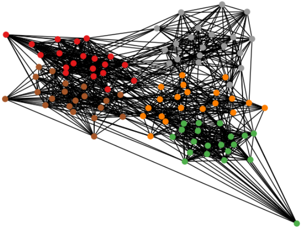
random edge w. induction



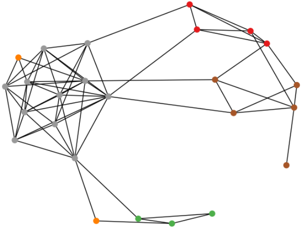
random edge w. partial induction

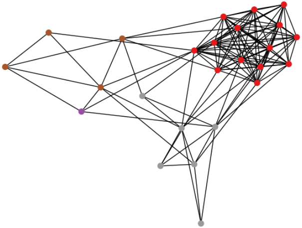
Samples: 25% of the nodes

original

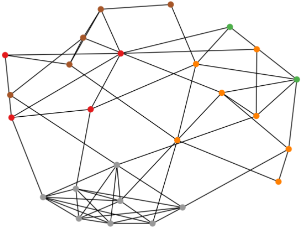


random walk





Metropolis Hastings



Shortest path

Example bias

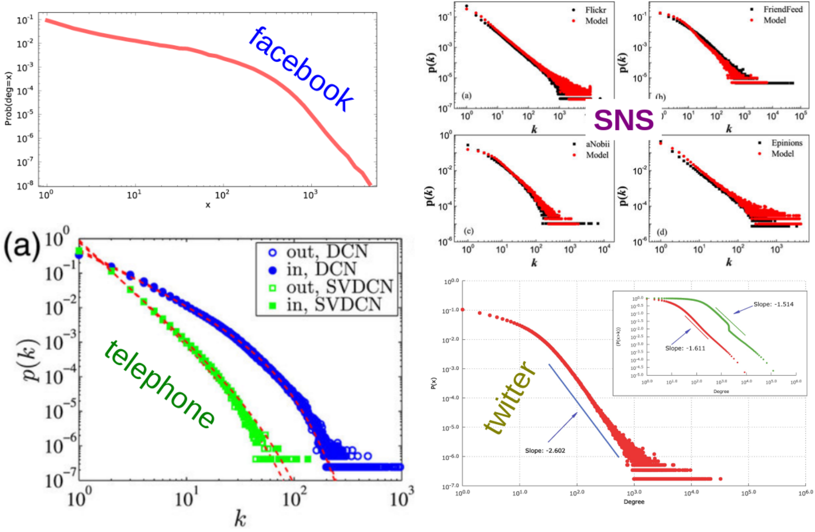
	BA	PPI	AS	arXiv
Degree Exponent	$\uparrow \uparrow \downarrow$	$\uparrow \uparrow =$	$= = \downarrow$	$\uparrow \uparrow \downarrow$
Average Path Length	$\uparrow \uparrow =$	$\uparrow \uparrow \downarrow$	$\uparrow \uparrow \downarrow$	$\uparrow \uparrow \downarrow$
Betweenness	$\uparrow \uparrow \downarrow$	$\uparrow \uparrow \downarrow$	$\uparrow \uparrow \downarrow$	$= = =$
Assortativity	$= = \downarrow$	$= = \downarrow$	$= = \downarrow$	$= = \downarrow$
Clustering Coefficient	$= = \uparrow$	$\uparrow \downarrow \uparrow$	$\downarrow \downarrow \uparrow$	$\downarrow \downarrow \downarrow$

Lee *et al* (2006): Entries indicate direction of bias for induced subgraph (red), incident subgraph (green), and snowball (blue) sampling.

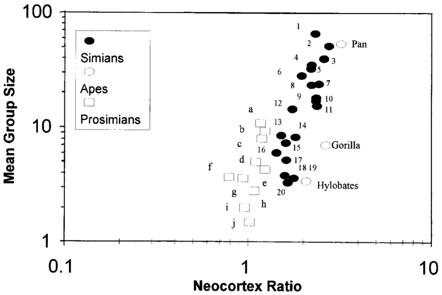
Sampling by ICT data

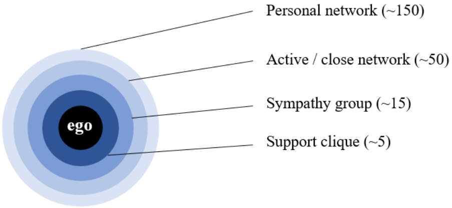
- ICT data: Samples society by a communication channel
- Knowledge is always partial
  - data is temporal
  - data displays part of the structure
- All sampling process alters the network structure.
- Main question: To what extent partial data can be use to describe the original system?

ICT data: degree distribution



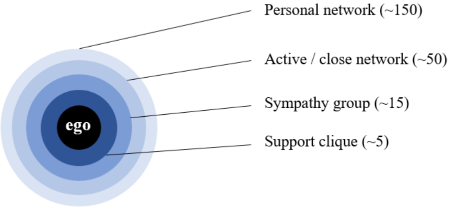
Dunbar number: 150

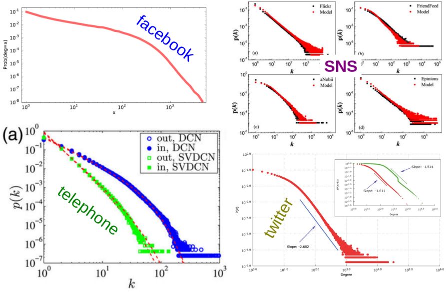




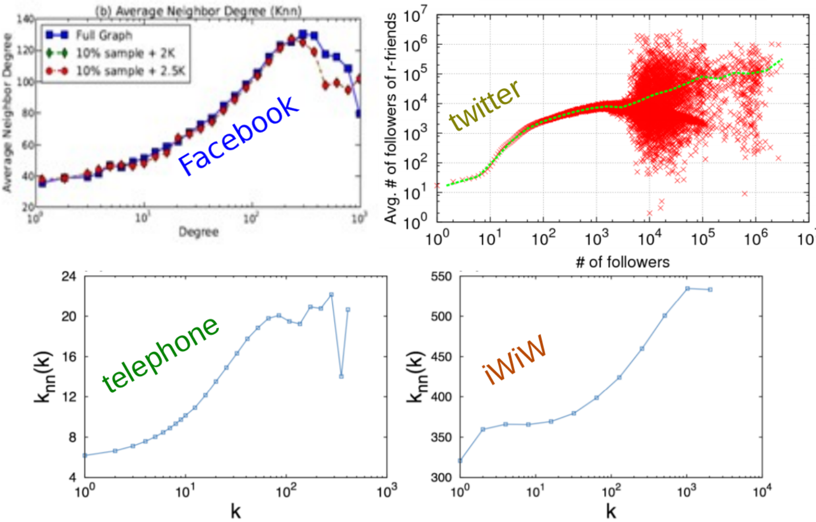
Dunbar number vs. ICT degree distribution

- Do we know anyone who has one single acquaintance?
- This must have been the most frequent case!



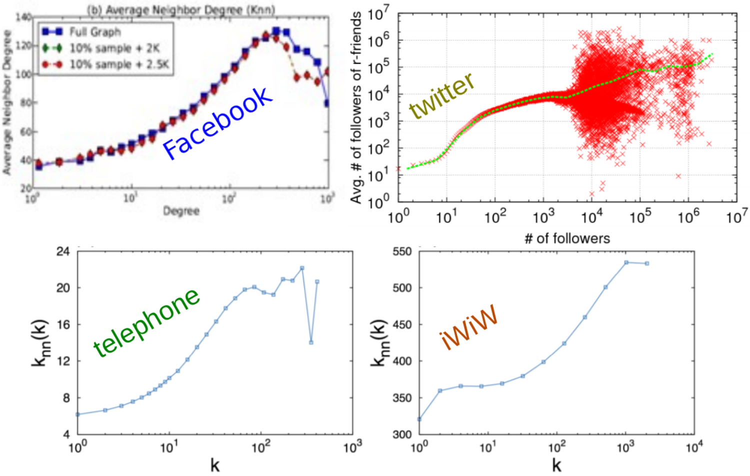


ICT data: assortativity

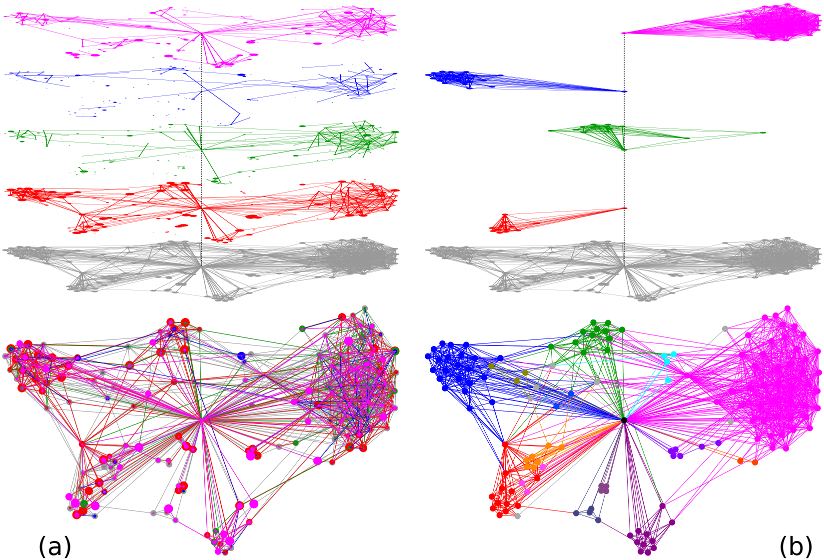


ICT data: assortativity

- Different system, similar curve!
- What do they show?



Social network and ICT data: Multiplex network

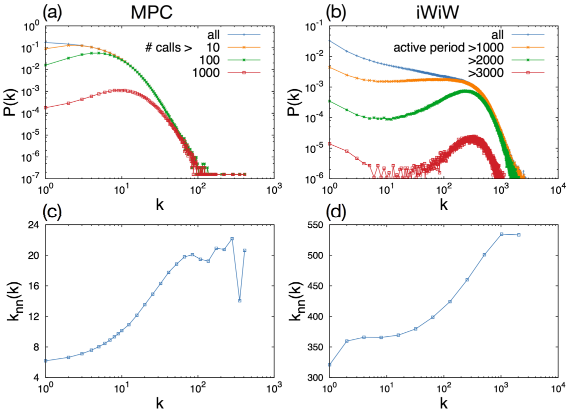


ICT data

- ICT data is always partial
- Most of the people do not live all their life in an online service (though we all know some who does)
- There is also a strong time factor (we need time to fully adapt a service)
- There is also personnel preference
- Certain communication channels are not apt for certain tasks

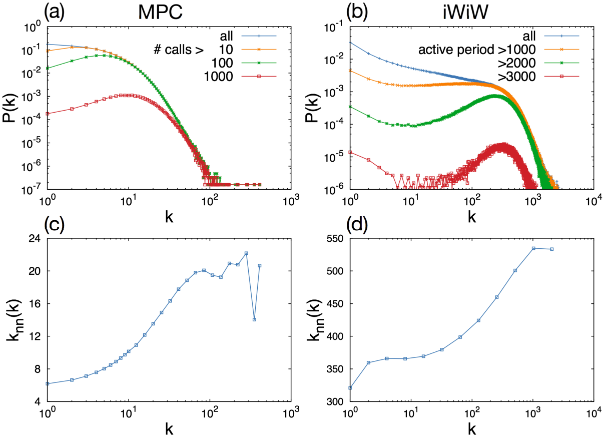
ICT data: Observations

- Degree distribution
  - It is always decreasing
  - Can it be reality?
- Assortativity
  - Increasing
  - Shape looks universal. Why?



ICT data: Observations

- Degree distribution
  - It is always decreasing
  - Can it be reality?
  - Remark that experienced/enthusiastic users have a peaked degree distribution



ICT data model

- Agents use the ICT systems to communicate
- Agents may use q different communication channel
- Each agent i has a personal preference f_i^α for channel α
- Agents i and j want to communicate, which channel to use?
  - One's favorite? Of course not! (I may write an email to my son and he will read in a week time, it is event worse if he tries to chat with me over Skype)
  - So we use the least uncomfortable:

$$\min_\alpha(f_i^\alpha, f_j^\alpha)$$

- If communication channel (layer) α is studied the probability of a link between users i and j is

$$p_{ij}^\alpha = \min(f_i^\alpha, f_j^\alpha)$$

- Let us drop α and focus on a single communication channel

ICT data model for a communication channel

- We start from a surrogate network (can be anything)
- Each agent i has a personal preference f_i for the given channel
- f_i is taken from a decreasing probability distribution e.g.

$$P(f) = \frac{1}{f_0} e^{-f/f_0}$$

- Links between agents i and j are kept with probability

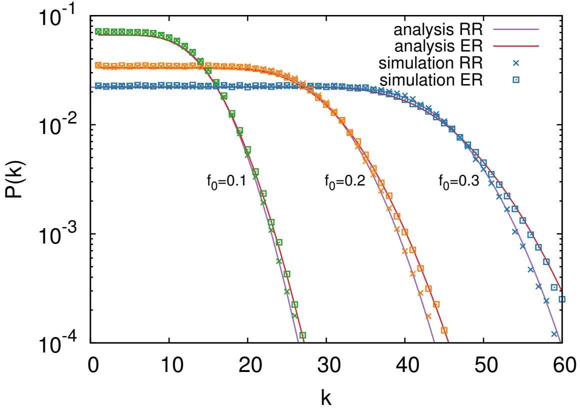
$$p_{ij} = \min(f_i, f_j)$$

ICT data model for a communication channel

- Analytic solution:

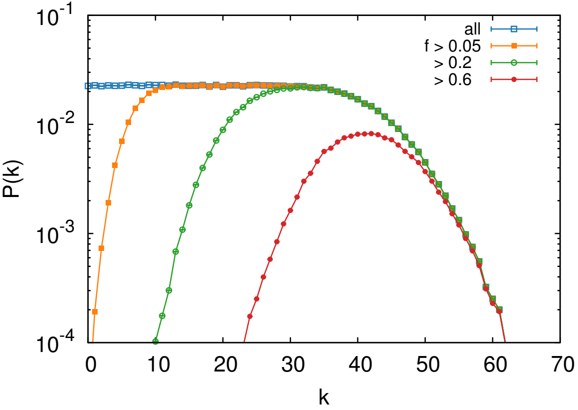
$$P(k) = \sum_{k'=0}^{\infty} P_0(k') \frac{1}{f_0(k'+1)} I_{\left(\frac{f_0}{1-f_0}\right)}(k+1, k'-k+1)$$

where $I_x(a, b)$ is the regularized beta function.

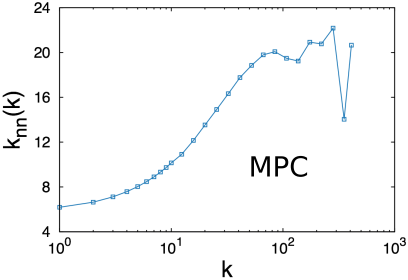


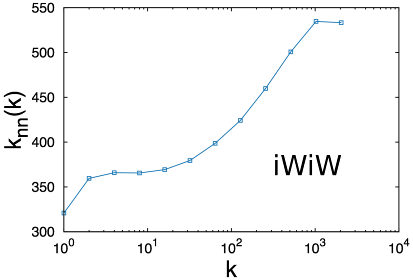
ICT data model: degree distribution

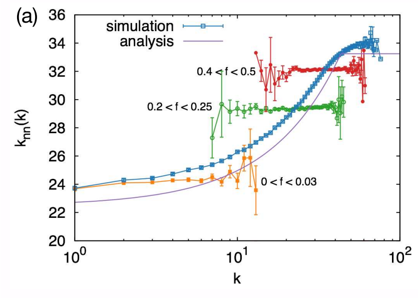
- Degree distribution changes from peaked to a monotonously decreasing one
- Devoted users have peaked degree distribution
- Surrogate network ER with $\langle k \rangle$=150



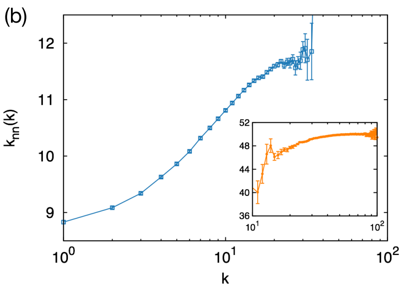
ICT data model: assortativity







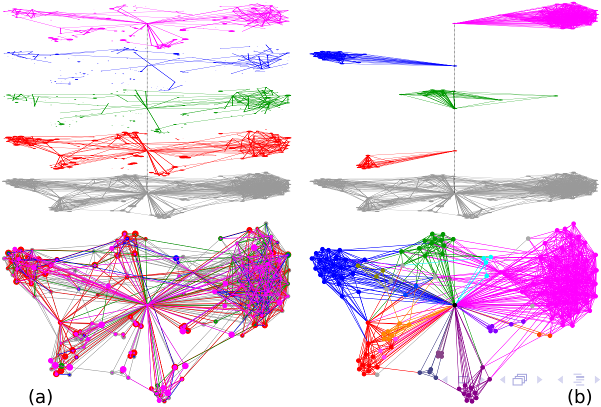
Sampled Erdős-Rényi



Sampled Social network

ICT data model: message

- ICT data is a biased sampling of the original network
- Properties may be results of the sampling/link selection process
- Original features may be totally invisible
- Experienced users in data are more similar to the original network



(a) (b)