Empirical Distribution of the Degree H-Index

Emma S. Spiro

Department of Sociology
University of California - Irvine

April 24, 2009

Task Outline and Approach

QUESTIONS:

1. What is the empirical distribution of h-index in real-world social networks?
2. What is the worst case scaling of the h-index as network size increases?

Task Outline and Approach

QUESTIONS:

1. What is the empirical distribution of h-index in real-world social networks?
2. What is the worst case scaling of the h-index as network size increases?

APPROACH

1. Sample real-world social networks
2. Look at the sampling distribution of h-indices in this population of networks.
3. Use standard regression techniques to approximate a bound on the scaling of the h-index in real-world social networks.

Problems and Questions

- Our questions are very reasonable, but not currently explored in the field.
- Some past work: (Faust and Skvoretz, 2002), (Butts, 2001), (Davis and Leinhardt, 1972)
- These research questions are relevant to MURI projects, namely to provide support for CS algorithms.
- More questions:
  - Are current ad hoc approaches to this problem appropriate?
  - What is the population of networks from which to sample?
  - What strategies are available for sampling real-world networks?
  - Can we approximate the scaling of network statistics?
  - Can we classify or group networks by those with/without certain properties?

Finding a Representative Sample of Real-world Networks

- Current strategy: convenience sampling - what do we have?
    - UCI Network Data Repository
    - UCINET
    - Pajek datasets
    - Population studies: AddHealth, Urban Communes Data Set
- Can we define a population of typical networks?
- What is a representative or typical social network?
- What is the best method for sampling typical networks?

H-Index Scaling

- 136 network data sets from UCINET, Pajek, and UCI Network Data Repository
- Chosen to include a range of network sizes.

	min.	median	mean	max.
network size (n)	10	67	535.3	10616
h-index (h)	2	12	19.08	116
$\log n$	2.303	4.204	4.589	9.270
$\log h$	0.6931	2.4849	2.6150	4.7536
$\log h / \log n$	0.2014	0.6166	0.6006	1.0000

Table: Summary statistics for real-world network data

H-Index Distribution

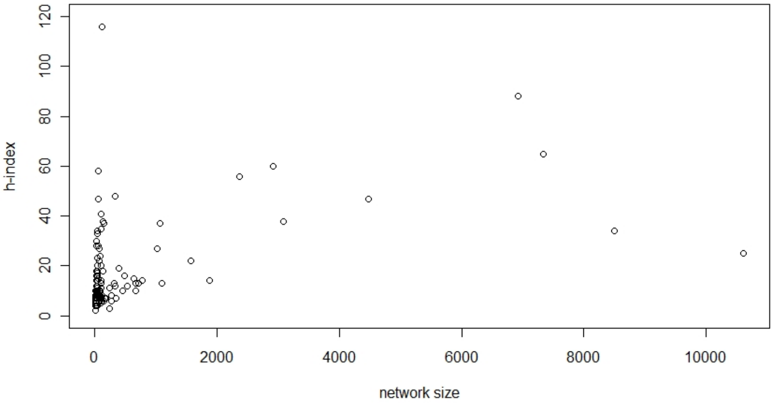

Figure: Scatter plot of *h*-index and network size

H-Index Distribution

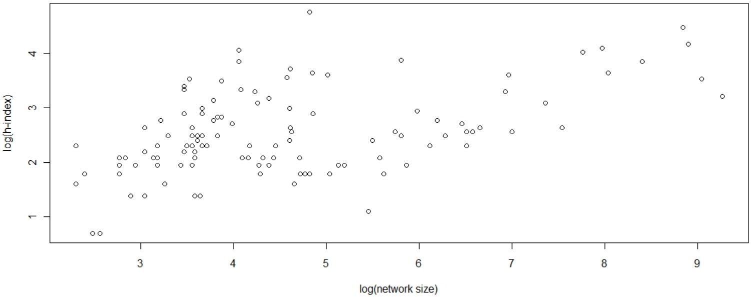

Figure: Scatter plot of *h*-index and network size

H-Index Distribution - Social/Non-social Grouping

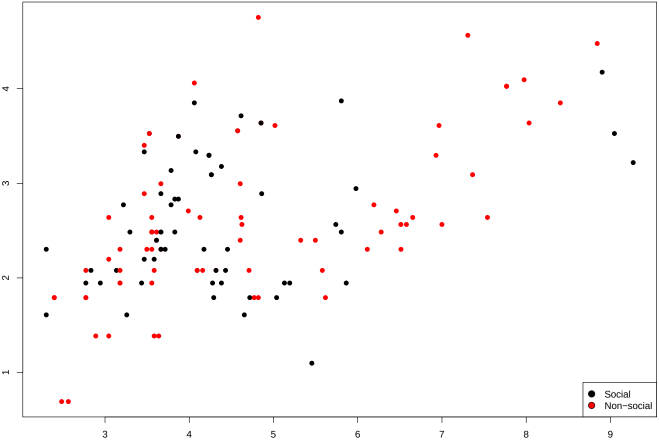

Figure: Scatter plot of *h*-index and network size

H-Index Distribution - Classification

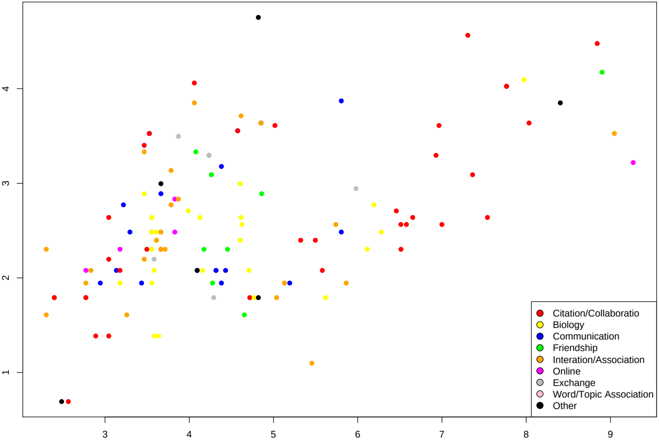

Figure: Scatter plot of *h*-index and network size

H-Index Distribution - Digraph/Graph Grouping

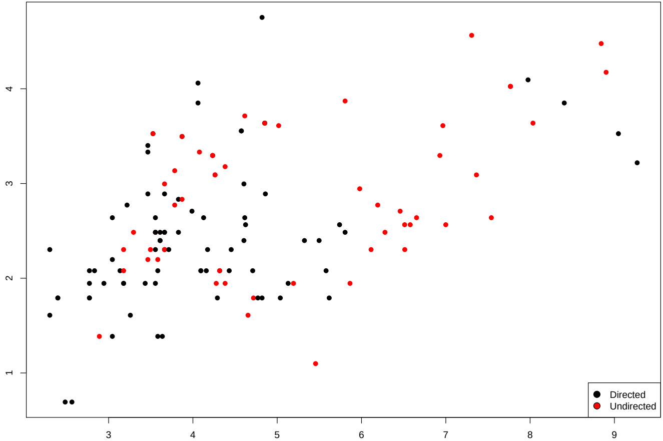

Figure: Scatter plot of *h*-index and network size

Latent Clusters of Networks

- We observe two clear clouds of data points in the empirical distribution.
- However, investigation do not yield any clear reason for the two clusters.
- We will use standard clustering algorithms to separate the network data sets into two classes.
- Clustering gives conservative estimates on scaling of the h-index with size.

Approximating Scaling

- Can we approximate the scaling of network statistics?
- What statistical approaches are appropriate for this problem?
  - Standard regression - approximation of the mean - not quite what we are interested in determining.
  - Quantile regression - quantiles - might be a better way to get at the scaling of network statistics.

Quantile Regression for H-Index Scaling

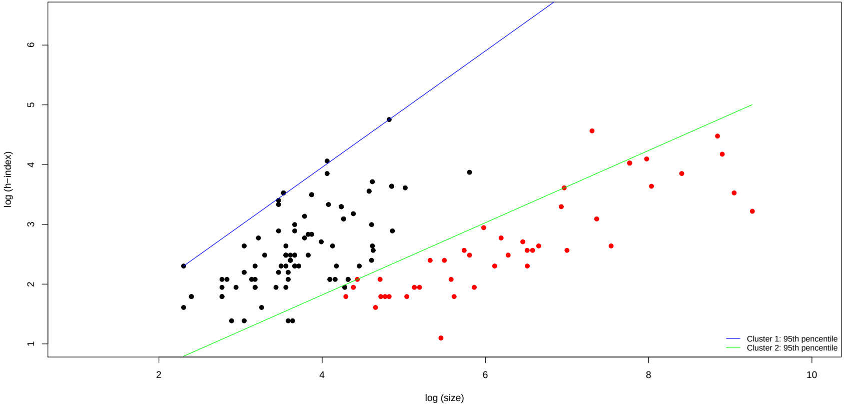

Figure: H-index scaling using quantile regression fits

Quantile Regression Results

Cluster	Intercept β_0	Slope β_1	df
1	0.0609 (-0.964, 2.581)	0.9735 (0.231, 1.266)	92
2	-0.598 (-1.938, 5.248)	0.604 (0.44712, 0.847)	44

Table: Coefficients for quantile regression lines

Cluster	log-like	AIC	BIC
1	-109.345	222.691	227.734
2	-41.071	86.143	89.712

Table: Goodness of fit measures for quantile regression lines

Problems and Questions - Future Research

- What strategies are available for sampling real-world networks?
- What methods can be used to approximate the scaling of network statistics?
- Is there a principled way to classify sets of networks?