

# Characterizing Influence in Google Buzz

CS 145 Final Presentation

Dallin Akagi  
Rishi Chandy  
Anthony Chong  
Manuel Lagang  
Jonathan Krause

# Outline

- Motivation & Background
- Measuring Social Influence
- Google Buzz
- Goals
- Approach
- Current Status & Results
- Prototype

# Measuring Social Influence

- Social influence is related to social power
- Follower graph model
- Standard Metrics
  - Indegree
  - Betweenness
  - H-index
  - Diffusion models (viral marketing)

# Google buzz



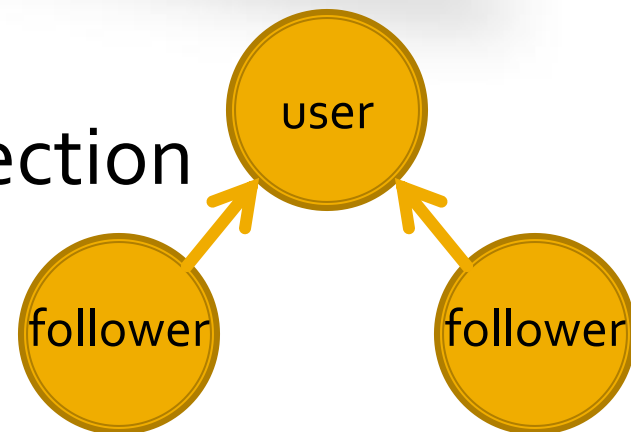
=



+



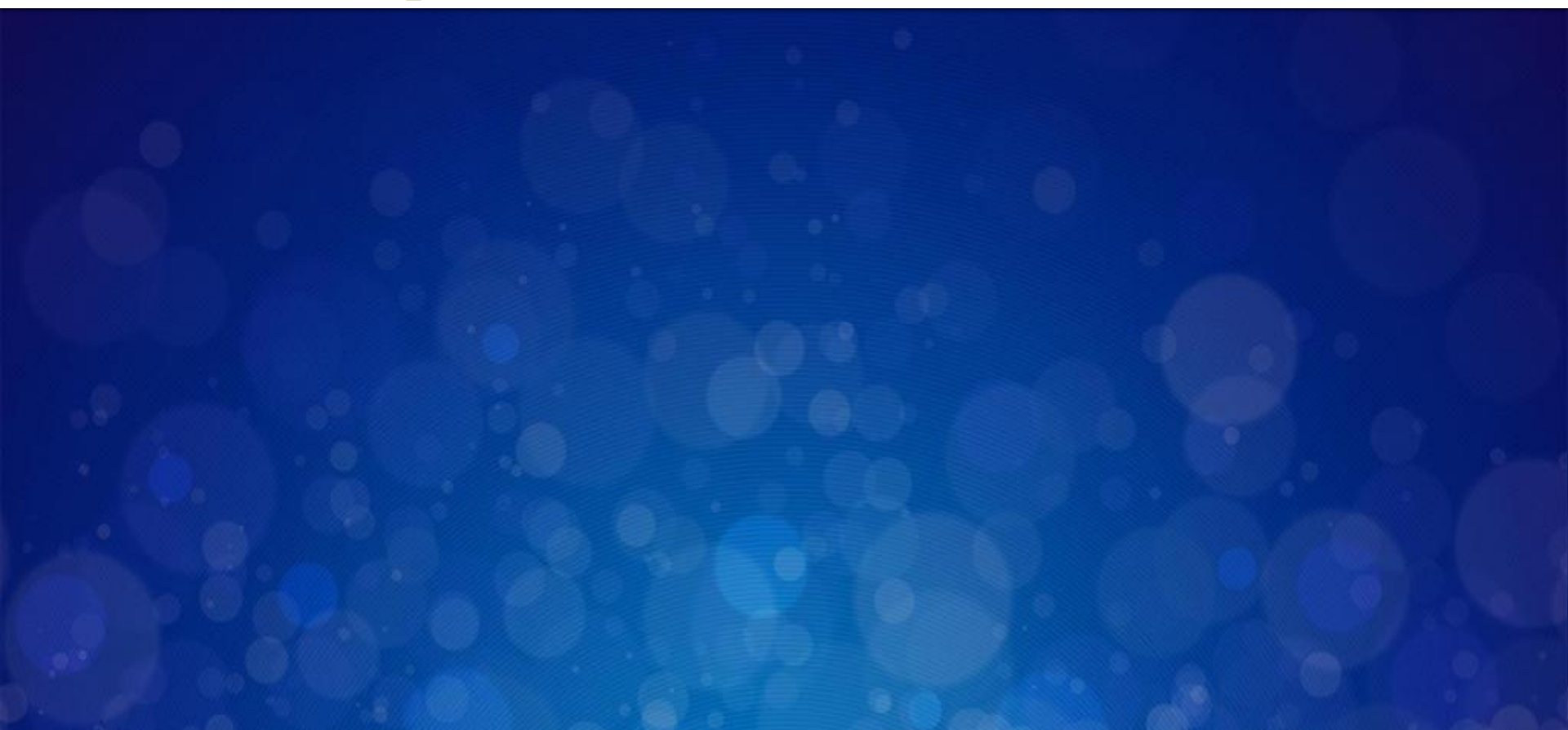
- Initialized with Gmail contacts
- Influence flow: reverse edge direction
  - Except comments and likes



# Goals

- Mid-term
  - Scrape Google Buzz
    - Friends, comments, and likes
  - Implement web application
    - Friend recommendation based on influence
- Final
  - Develop and **evaluate** influence metrics
  - User analytics (“resonance”)

# Sample Bias



# Sample Bias

- Graph sampling techniques create sampling bias based on node degree
  - Can we characterize sample bias from our sampling technique?
  - Can we correct for this bias?
- Kurant, Markopoulou, Thiran (2010)

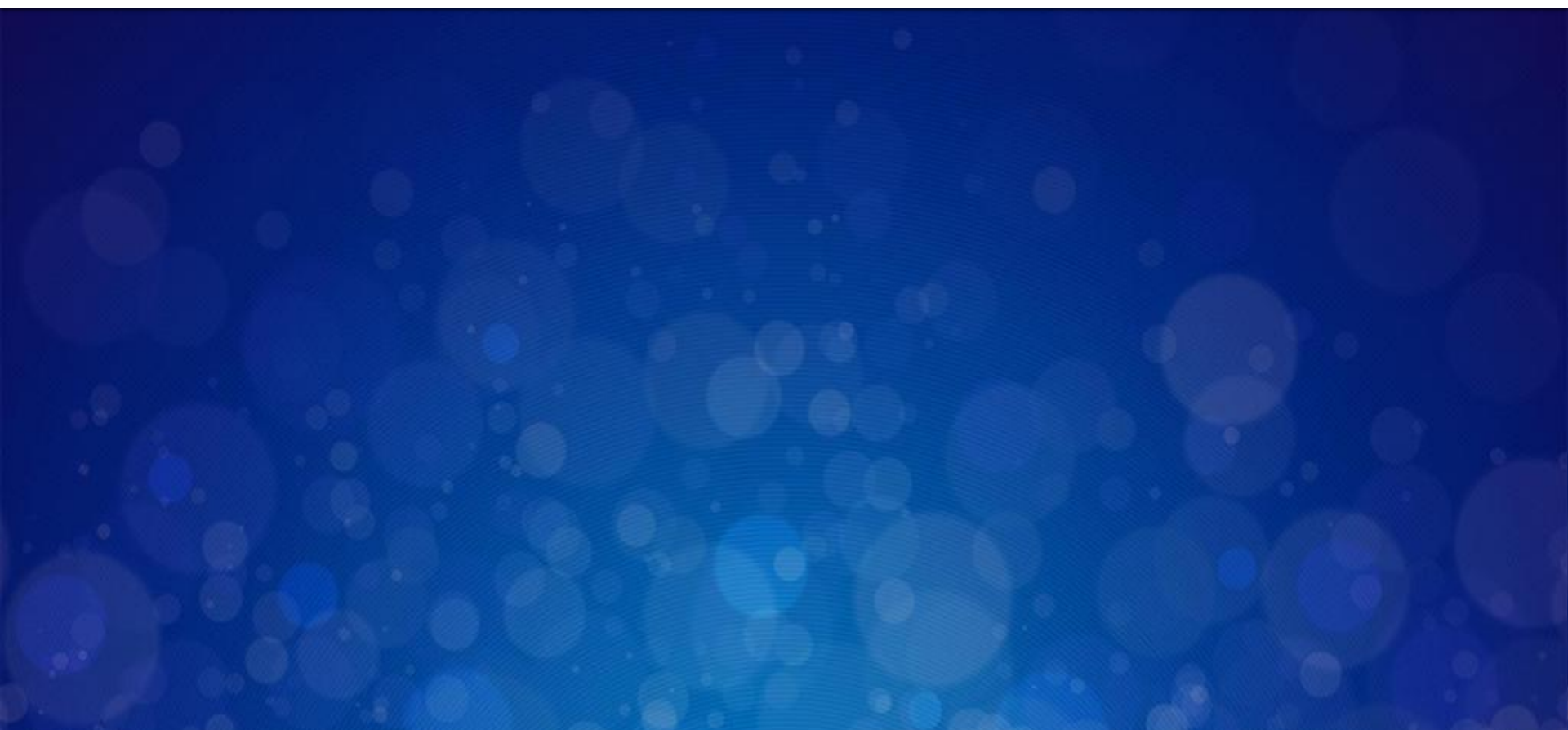
# Sample Bias (cont'd)

- Consider the family of Random Graphs given a degree distribution  $p_k$  denoted  $PW(p_k)$
- Degree distribution we observe ( $q_k$ ) is biased towards high-degree nodes.
- Bound bias by:

$$\hat{p}_k = \frac{\hat{q}_k}{1 - (1-t(f))^k} \cdot \left( \sum_l \frac{\hat{q}_l}{1 - (1-t(f))^l} \right)^{-1}$$

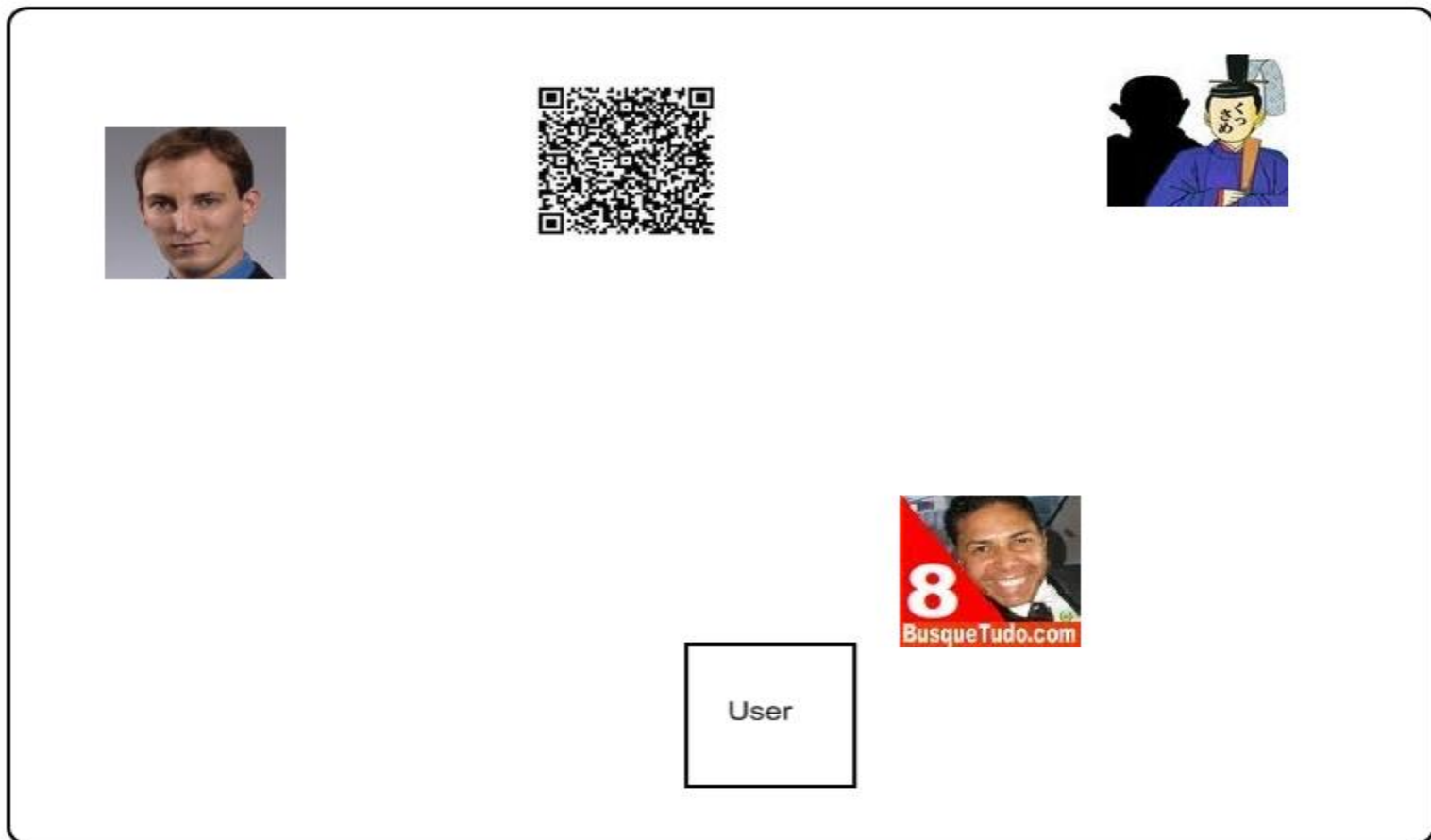
Unfortunately  $t(f)$  requires,  $p_k$  however, we can iteratively solve for this. Here,  $f$  is the fraction of covered nodes

# Personalization



# Personalizing Recommendations

- Global influence does not matter





# First attempt

- Limit to “nearby” globally influential people
- Problems:
  - “nearby” is arbitrary
  - no local graph structure
  - dominated by a few top global nodes

# Another idea

- Consider only a local subgraph
- Issues still present
  - Choice of local subgraph is arbitrary
  - Information is thrown away
  - Many metrics are inherently local





# What does this achieve?

- Measures influence on nodes local to user
- Uses whole graph structure
- “Local” is still arbitrary

# Metric Comparisons

# Metric Comparisons

- Kendall's Tau:

$$\tau = \frac{\sum_{i,j} [i, j \text{ in same order}] - \sum_{i,j} [i, j \text{ in different order}]}{\frac{1}{2}n(n-1)}$$

- In  $[-1,1]$ , 1 is perfect correlation
- Linear Regression: faster, less meaningful
  - Use  $R^2$

# Kendall's Tau (Buzz)

	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk
Hirsch Index	1	.2665	.8122	.2689	.2125	.0868
Ind. Cascade	.2665	1	.3645	.7823	.8140	.1382
In-Degree	.8122	.3645	1	.3645	.3090	.2411
In-Web <sub>2</sub>	.2689	.7823	.3645	1	.8349	.1056
In-Web <sub>3</sub>	.2125	.8140	.3090	.8349	1	.1021
Random Walk	.0868	.1382	.2411	.1056	.1021	1

# Linear Regression (First Degree)

	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk
Hirsch Index	1	.6799	.6159	.6852	.3807	.2720
Ind. Cascade	.6799	1	.2805	.8538	.7536	.1059
In-Degree	.6159	.2805	1	.3299	.1488	.5854
In-Web <sub>2</sub>	.6852	.8538	.3299	1	.7583	.1103
In-Web <sub>3</sub>	.3807	.7536	.1488	.7583	1	.0422
Random Walk	.2720	.1059	.5854	.1103	.0422	1

# Linear Regression (Second Degree)

	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk
Hirsch Index	1	.8454	.7475	.7479	.5105	.3057
Ind. Cascade	.8327	1	.4294	.8588	.8573	.1939
In-Degree	.8529	.4979	1	.5052	.2681	.5856
In-Web <sub>2</sub>	.8409	.8605	.6392	1	.8614	.2646
In-Web <sub>3</sub>	.6841	.8550	.3574	.9722	1	.1449
Random Walk	.3004	.1285	.6196	.1186	.0487	1

# Kendall's Tau (Stack Overflow)

	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk	Reputation
Hirsch Index	1	.3921	.6752	.3954	.3772	.2205	.0863
Ind. Cascade	.3921	1	.5964	.8719	.8742	.4958	.2749
In-Degree	.6752	.5964	1	.5976	.5796	.4675	.2118
In-Web <sub>2</sub>	.3954	.8719	.5976	1	.9299	.4868	.2616
In-Web <sub>3</sub>	.3772	.8742	.5796	.9299	1	.4843	.2597
Random Walk	.2205	.4958	.4675	.4868	.4843	1	.2597
Reputation	.0863	.2749	.2118	.2616	.2597	.2597	1

# Linear Regression (First Degree)

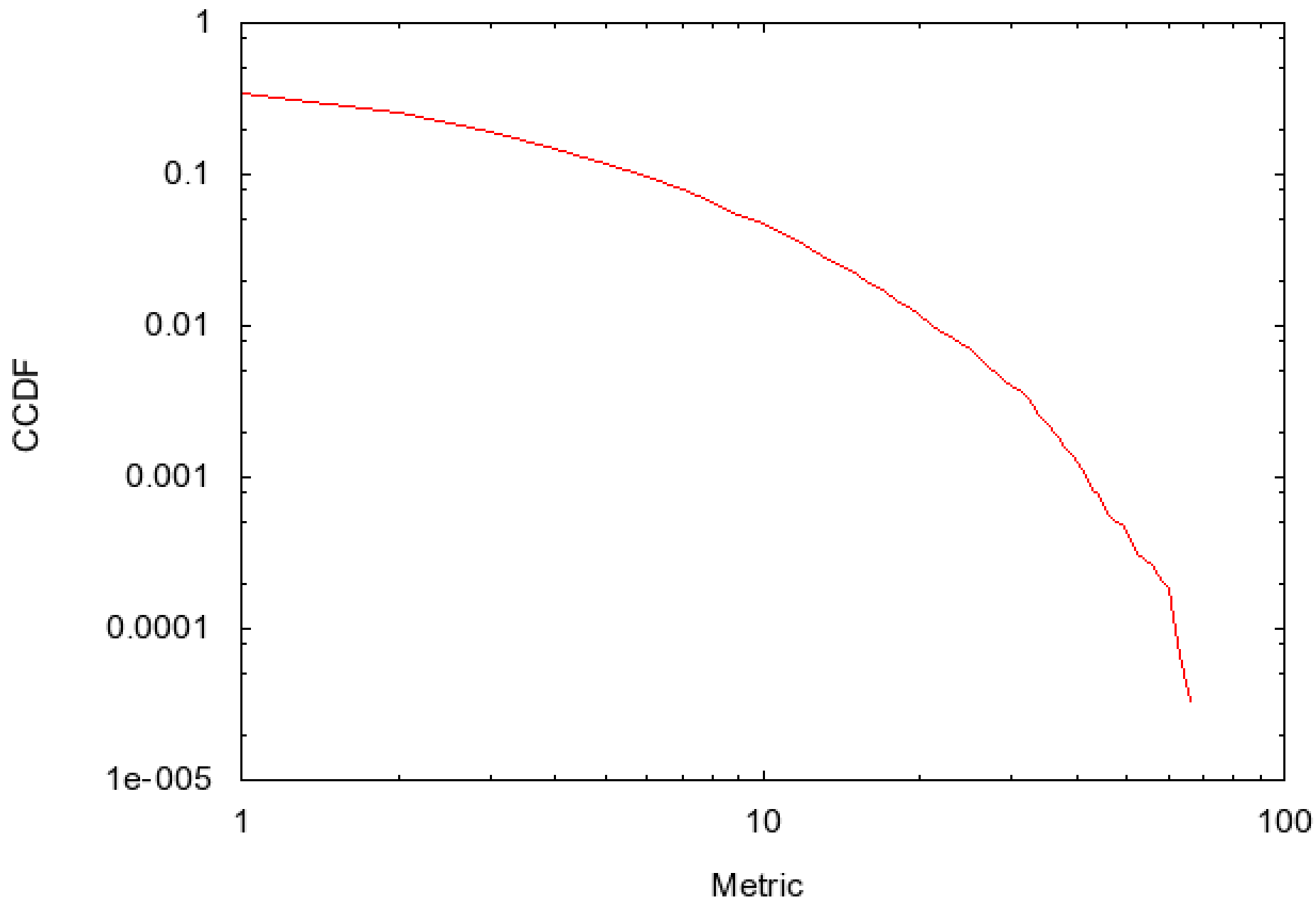
	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk	Reputation
Hirsch Index	1	.8742	.6213	.8121	.8204	.3751	.1246
Ind. Cascade	.8742	1	.3896	.6559	.8945	.2205	.1016
In-Degree	.6213	.3896	1	.7491	.4098	.7169	.0969
In-Web <sub>2</sub>	.8121	.6559	.7491	1	.7639	.4958	.1091
In-Web <sub>3</sub>	.8204	.8945	.4098	.7639	1	.2531	.0947
Random Walk	.3751	.2205	.7169	.4958	.2531	1	.0418
Reputation	.1246	.1016	.0969	.1091	.0947	.0418	1

# Linear Regression (Second Degree)

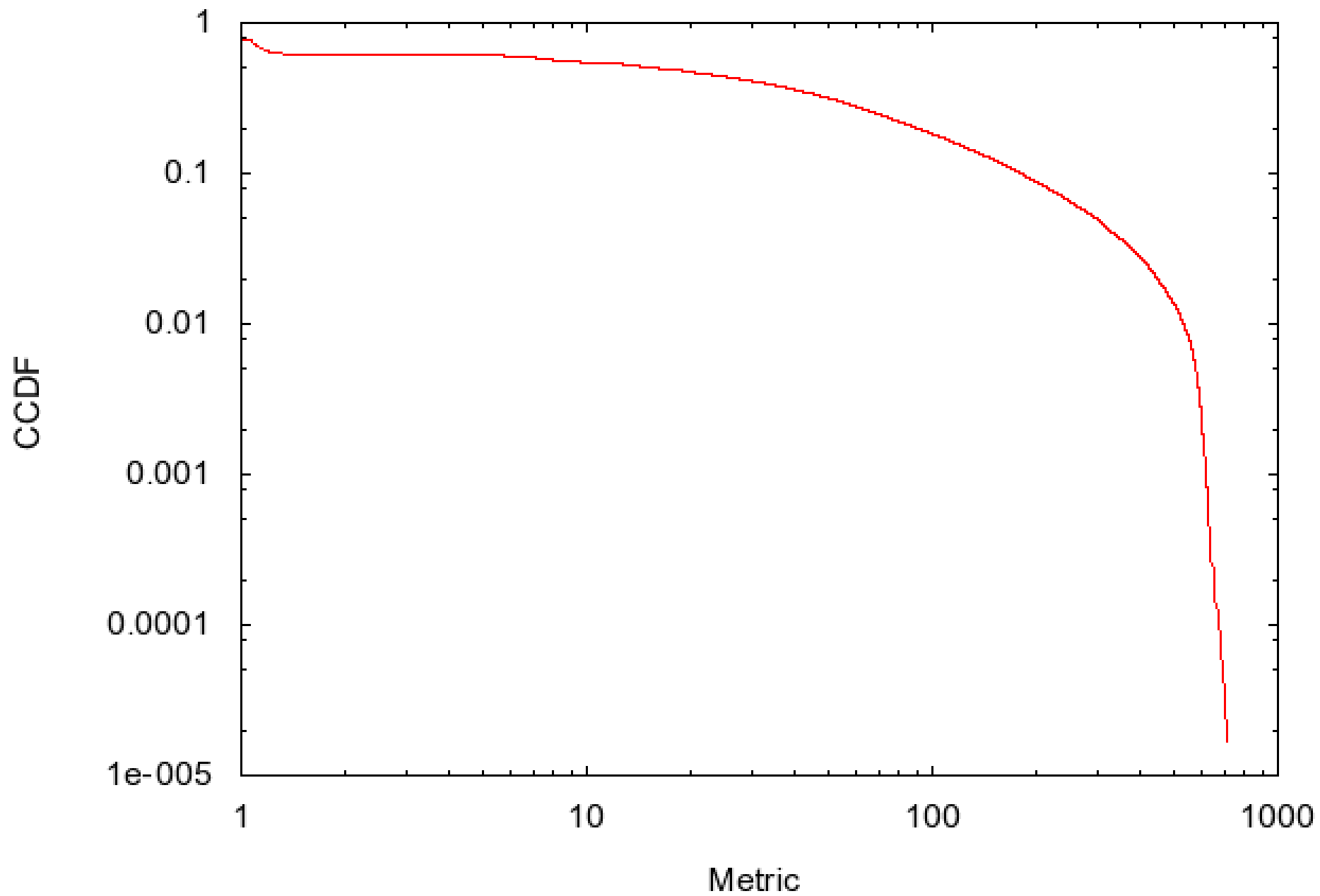
	Hirsch Index	Ind. Cascade	In-Degree	In-Web <sub>2</sub>	In-Web <sub>3</sub>	Random Walk	Reputation
Hirsch Index	1	.9235	.9036	.9012	.8393	.5963	.1291
Ind. Cascade	.9100	1	.5612	.7935	.8950	.3281	.1121
In-Degree	.8158	.6161	1	.8558	.6059	.7477	.1145
In-Web <sub>2</sub>	.8496	.8216	.8880	1	.9080	.6426	.1092
In-Web <sub>3</sub>	.8564	.8966	.6635	.9608	1	.4313	.1052
Random Walk	.5097	.3479	.7265	.6070	.3834	1	.0602
Reputation	.1620	.1459	.1028	.1311	.1328	.0452	1

**Now let's look at some  
CCDFs...**

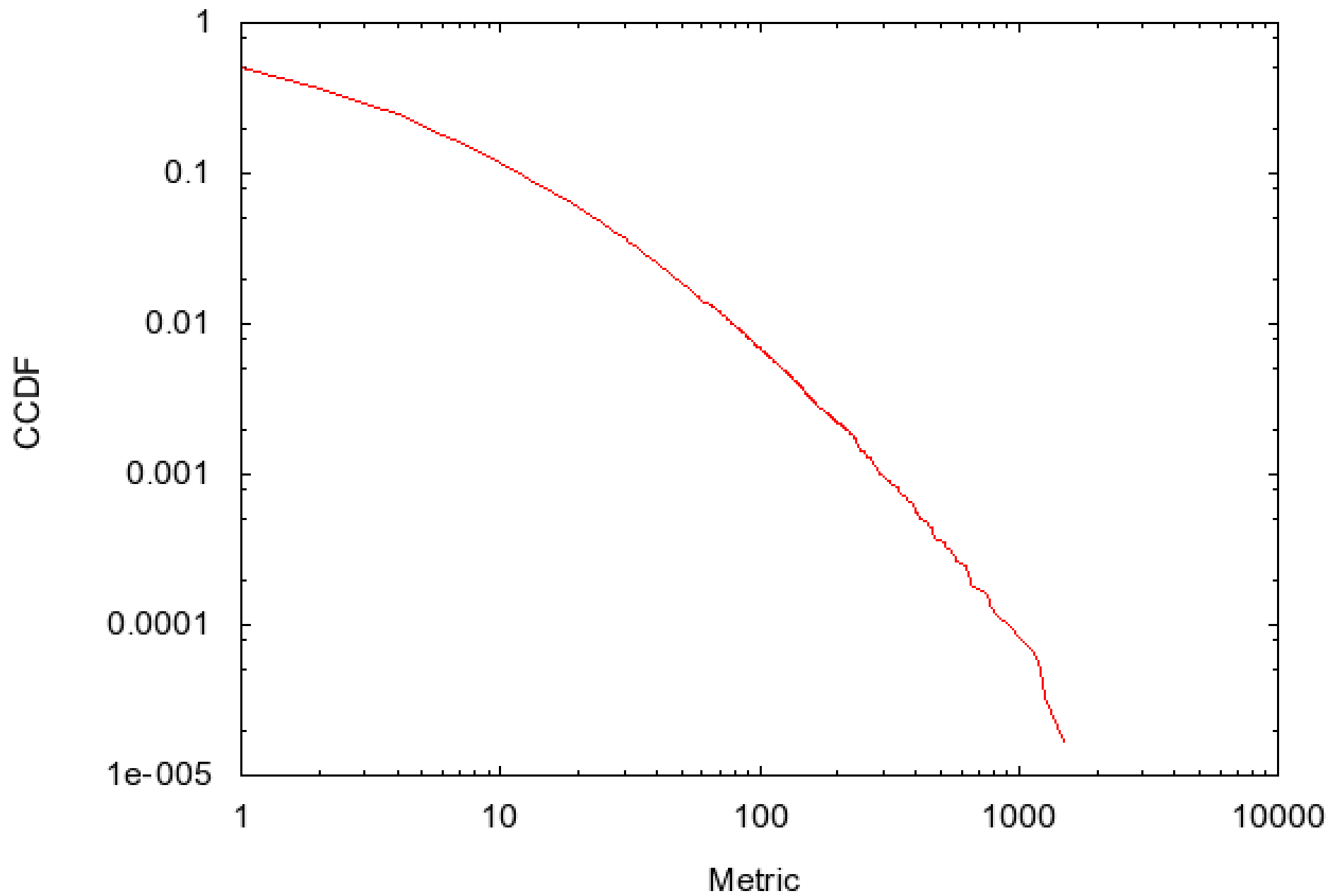
Hirsch Index CCDF, Stack Overflow



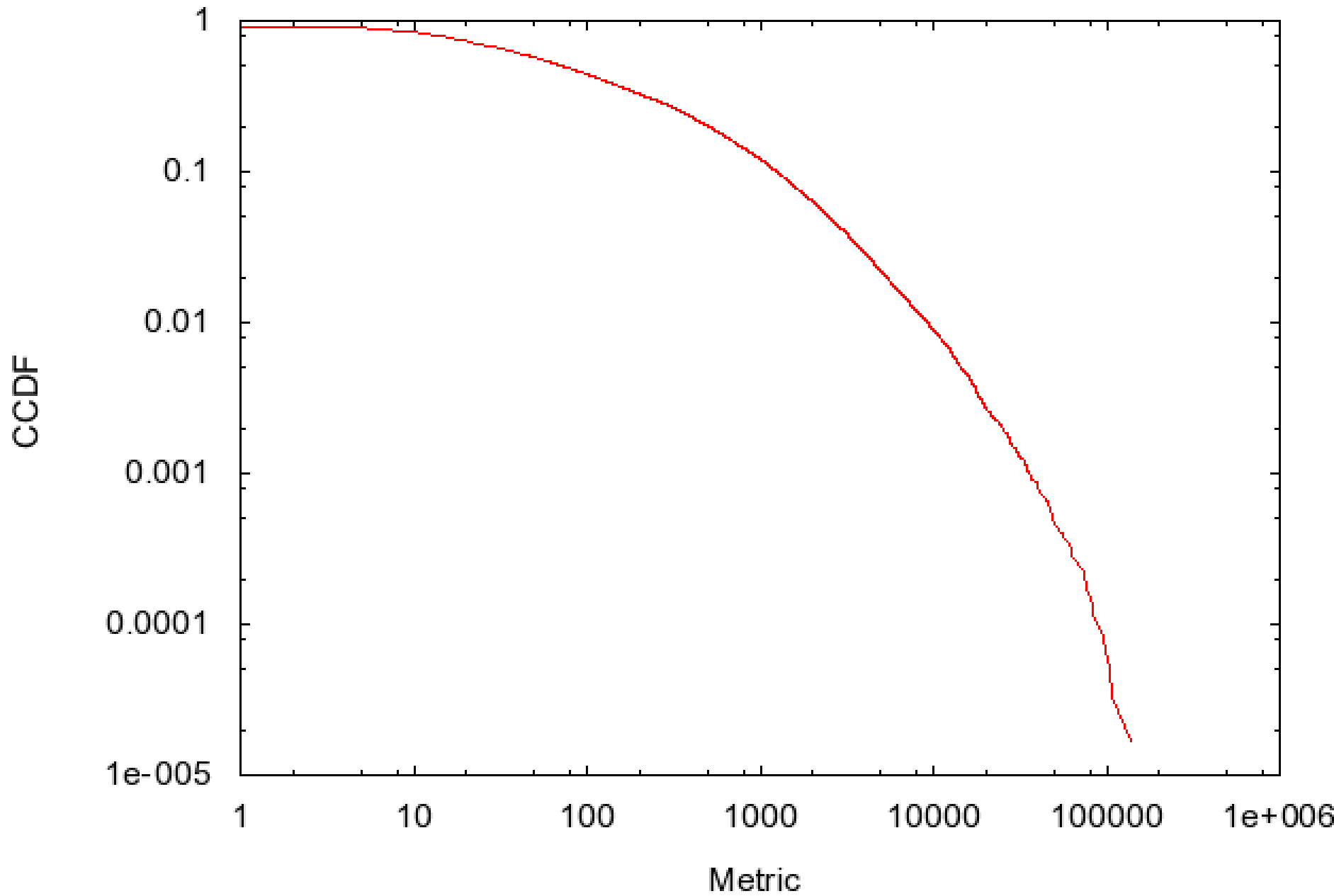
Independent Cascade CCDF, Stack Overflow



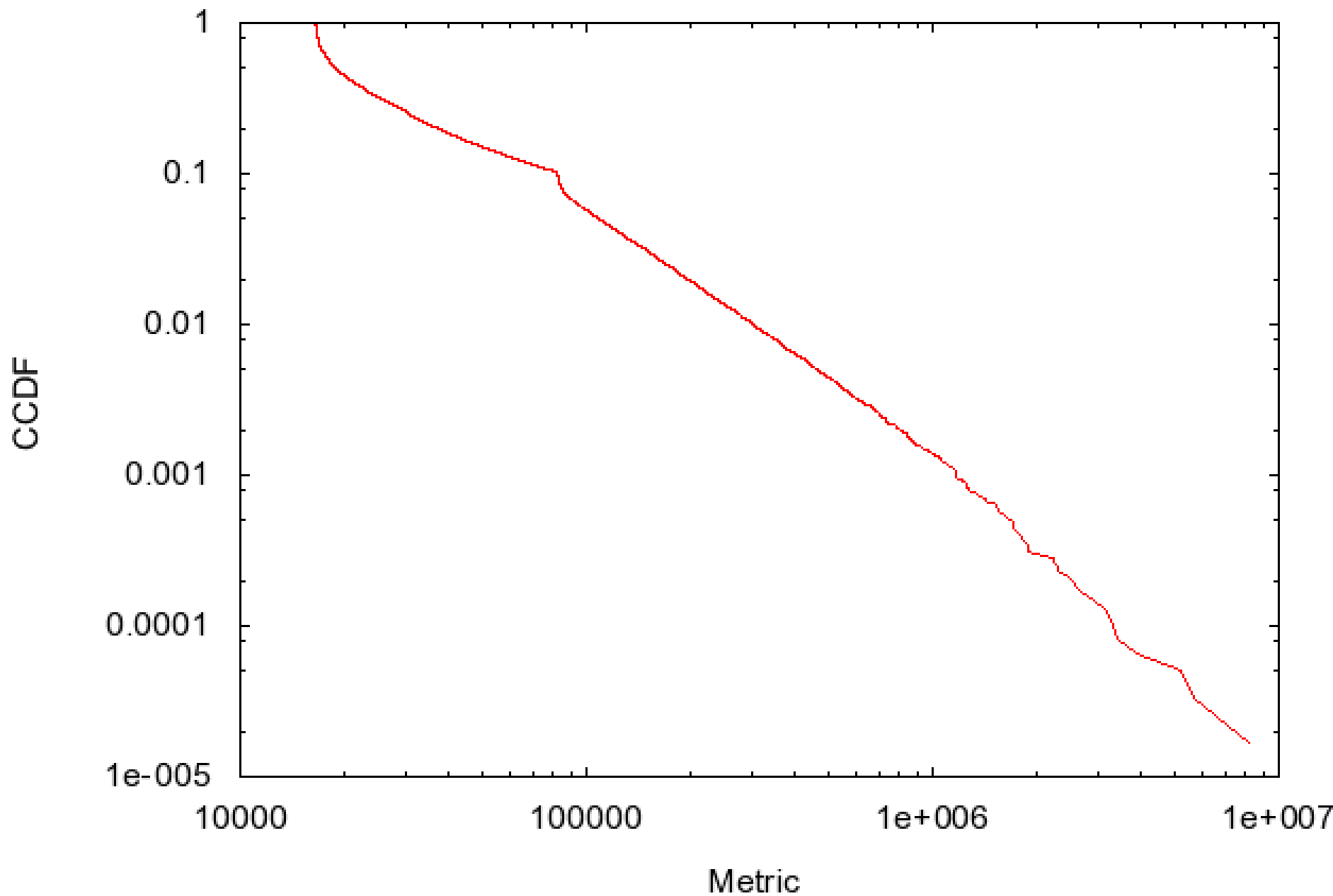
In-Degree CCDF, Stack Overflow



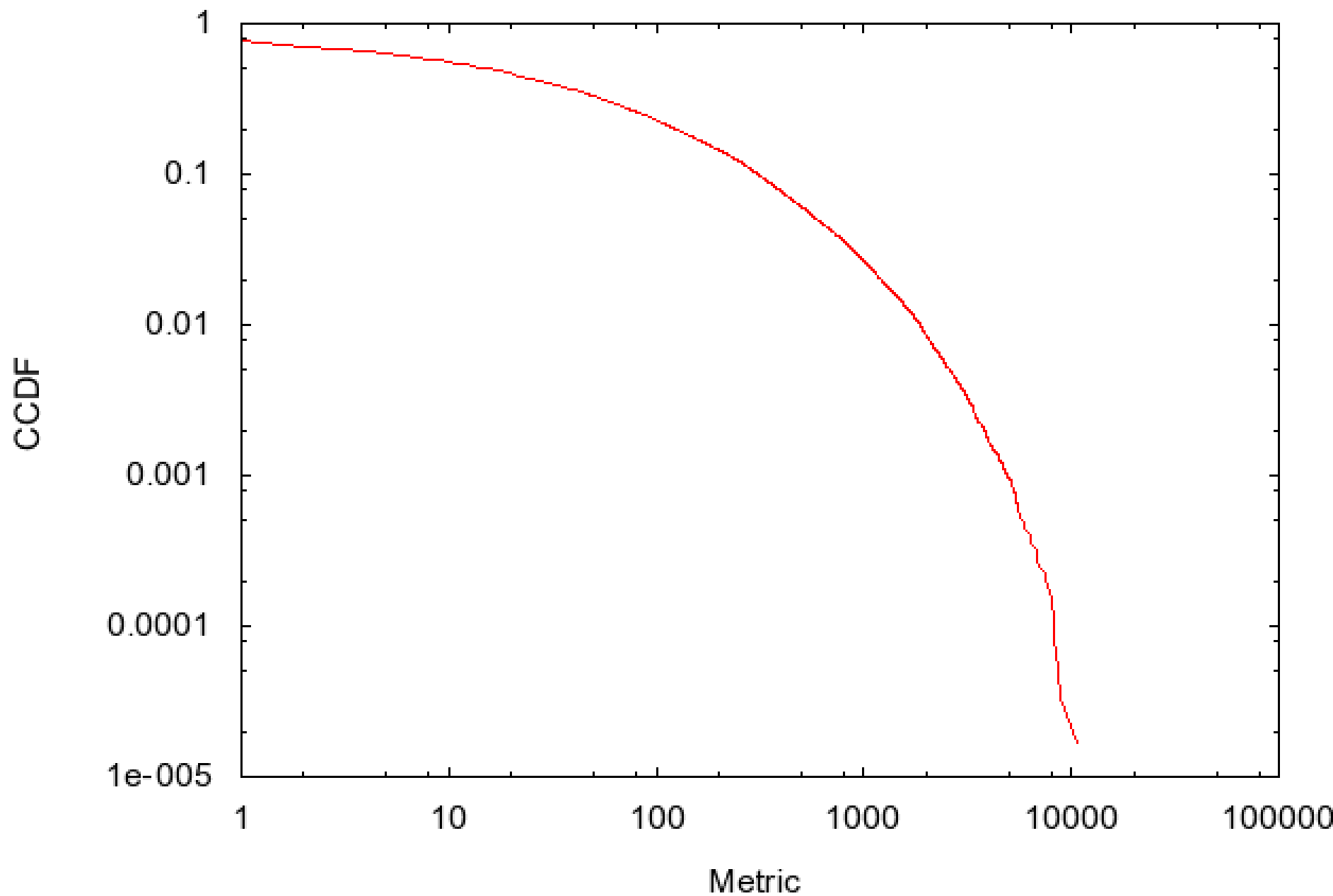
Reputation CCDF, Stack Overflow



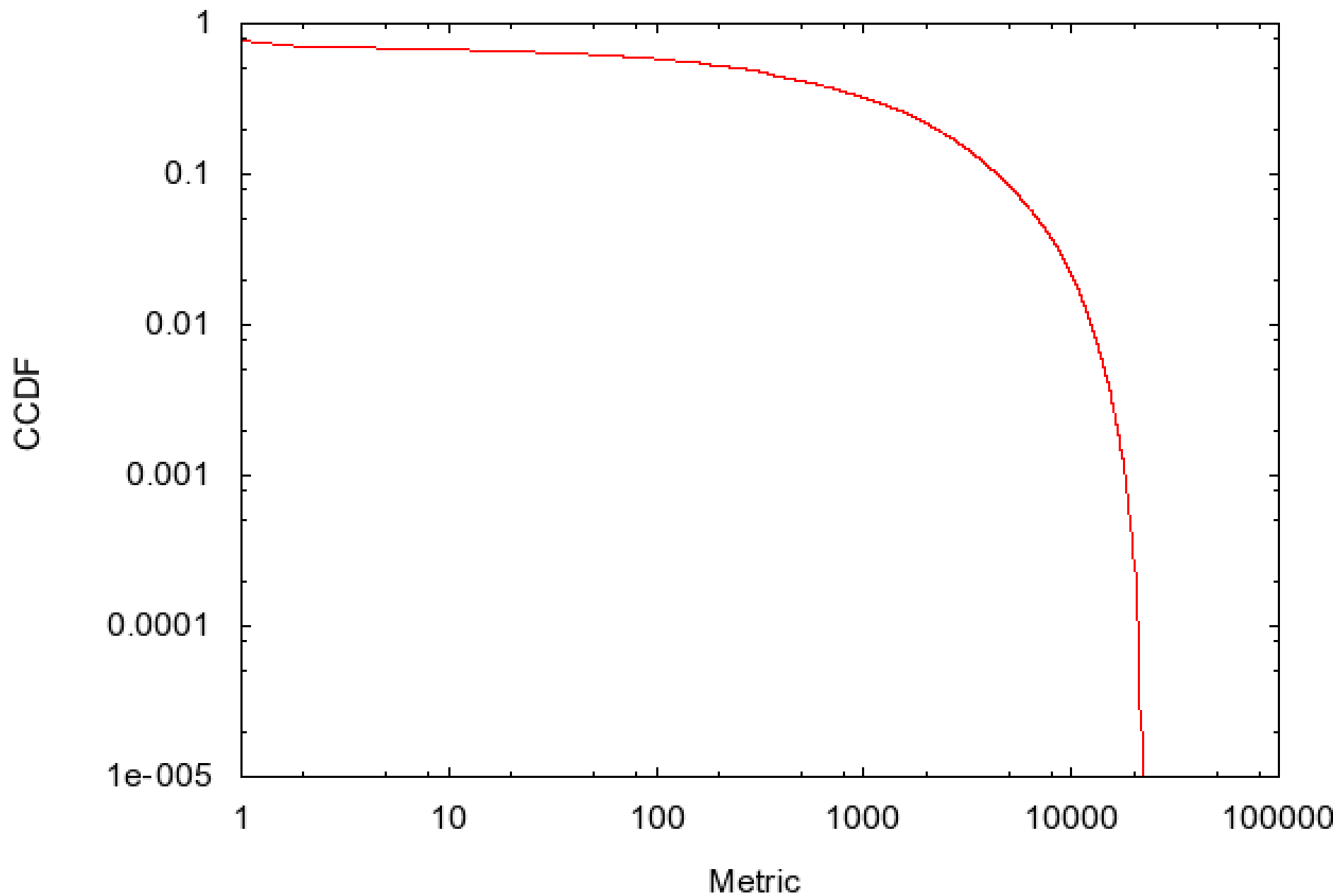
Random Walk (1B, .2) CCDF, Stack Overflow



In-Web\_2 CCDF, Stack Overflow



In-Web\_3 CCDF, Stack Overflow



# One Metric to Rule Them All

# Towards One Metric

Dwork et al, 2001

- Acquire an aggregate  $\sigma$  (we use Borda's method) of all orderings  $\tau_1, \tau_2 \dots$
- Do adjacency swaps on the aggregate which will lower Kendall distance (called local Kemenization)  
$$K(\sigma', \tau_1, \tau_2, \dots, \tau_k) < K(\sigma, \tau_1, \tau_2, \dots, \tau_k)$$
- When no more swaps are possible, have reached the (unique) local Kemenization
- Resultant aggregate is maximally consistent with original, satisfies Extended Condorcet Criterion

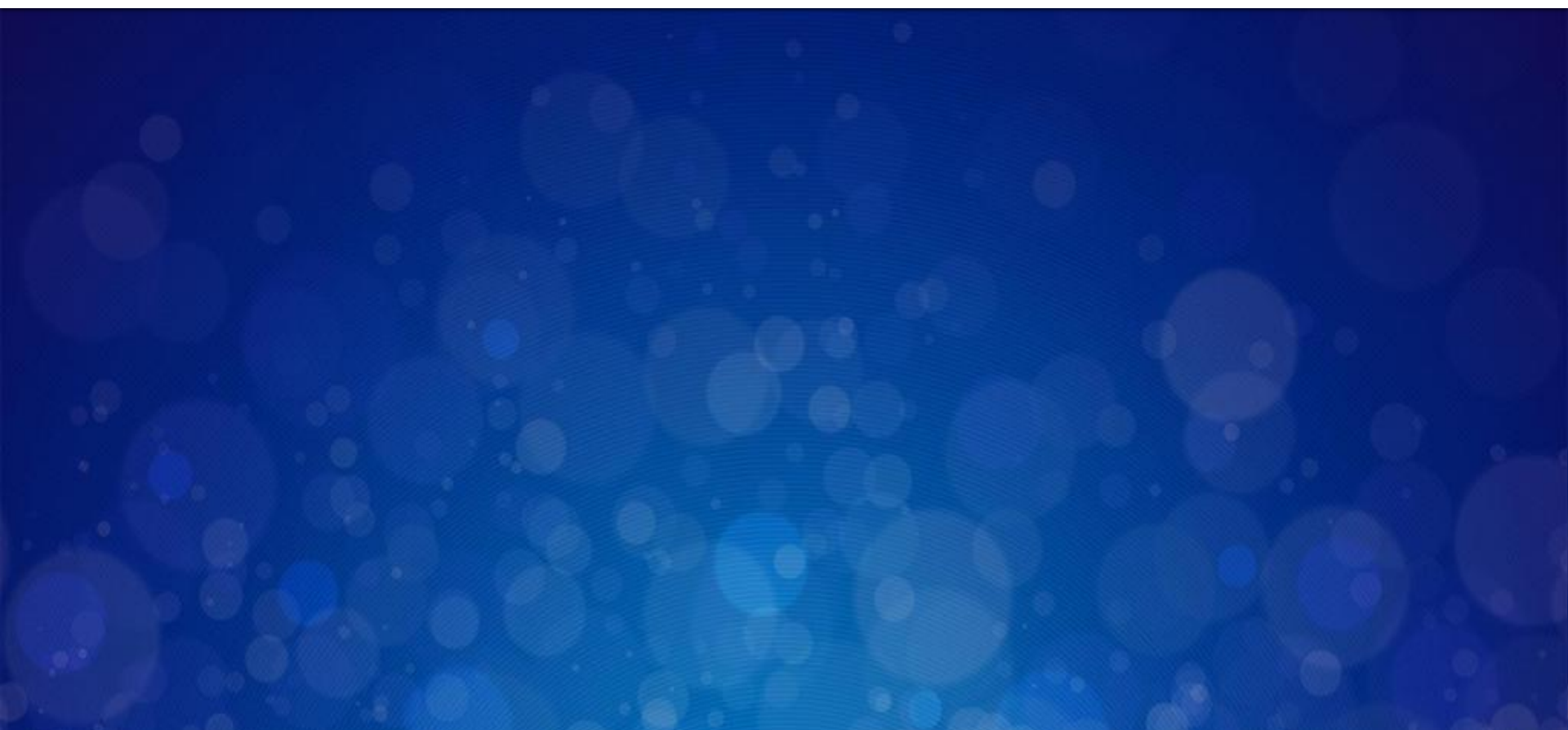
# Towards One Metric

- Result depends on initial quality of  $\sigma$
- Gives equal importance to all metrics (hence the importance of the above)
- ECC lends itself to combating spam
- Local Kemenization can be approximated by a polynomial time algorithm


# Refined Golf Score

- “Golf Score” = “Borda’s Method”
  - Sum of ranks (lower is better)
- Refined using Kendall Tau minimization (local Kemenization)


# Evaluation



# Evaluation: Stack Overflow

- Test Dataset:  **stackoverflow**
  - Technical Q&A website
  - Friend graph derived from “favorites”
  - Public “reputation” score

# Evaluation: Stack Overflow

- Test Dataset:  **stackoverflow**
  - Technical Q&A website
  - Friend graph derived from “favorites”
  - Public “reputation” score
- Results
  - No relationship among our metrics and “reputation”
  - “Favorites” is probably not a social relationship

# Prototype

