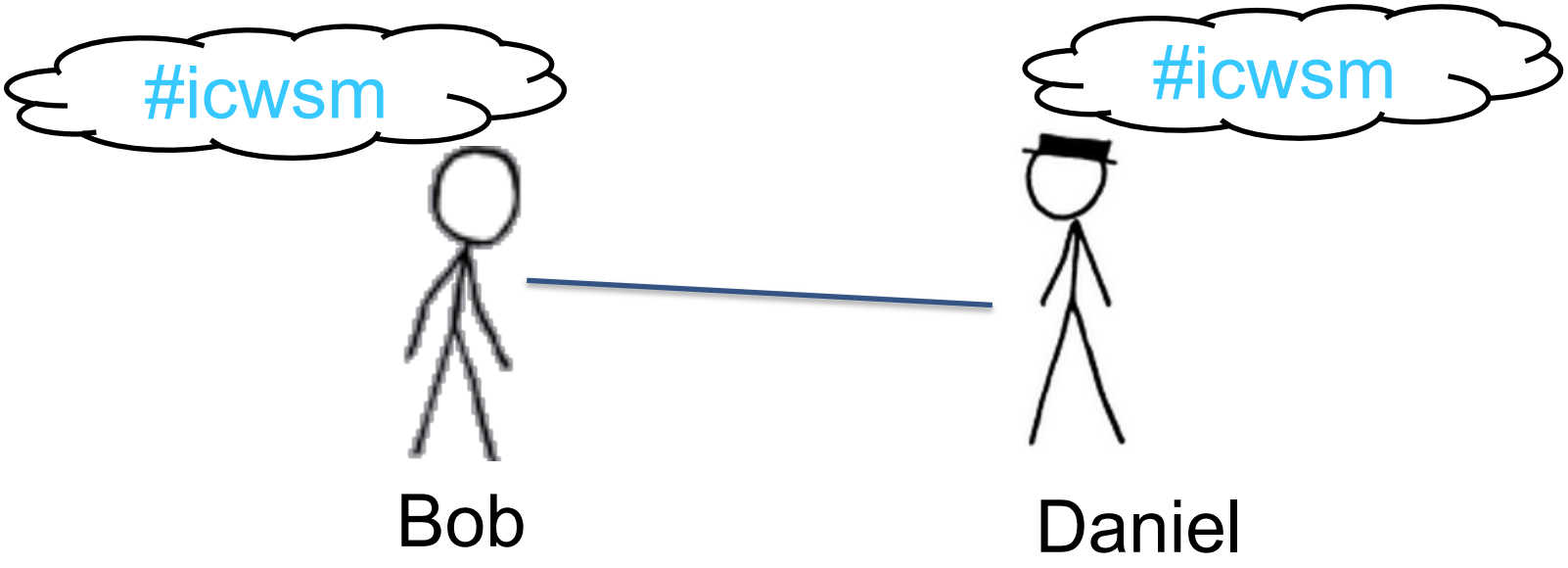
On the Interplay between Social and Topical Structure

Daniel M. Romero, *Chenhao Tan*, Johan Ugander
Northwestern University & Cornell University

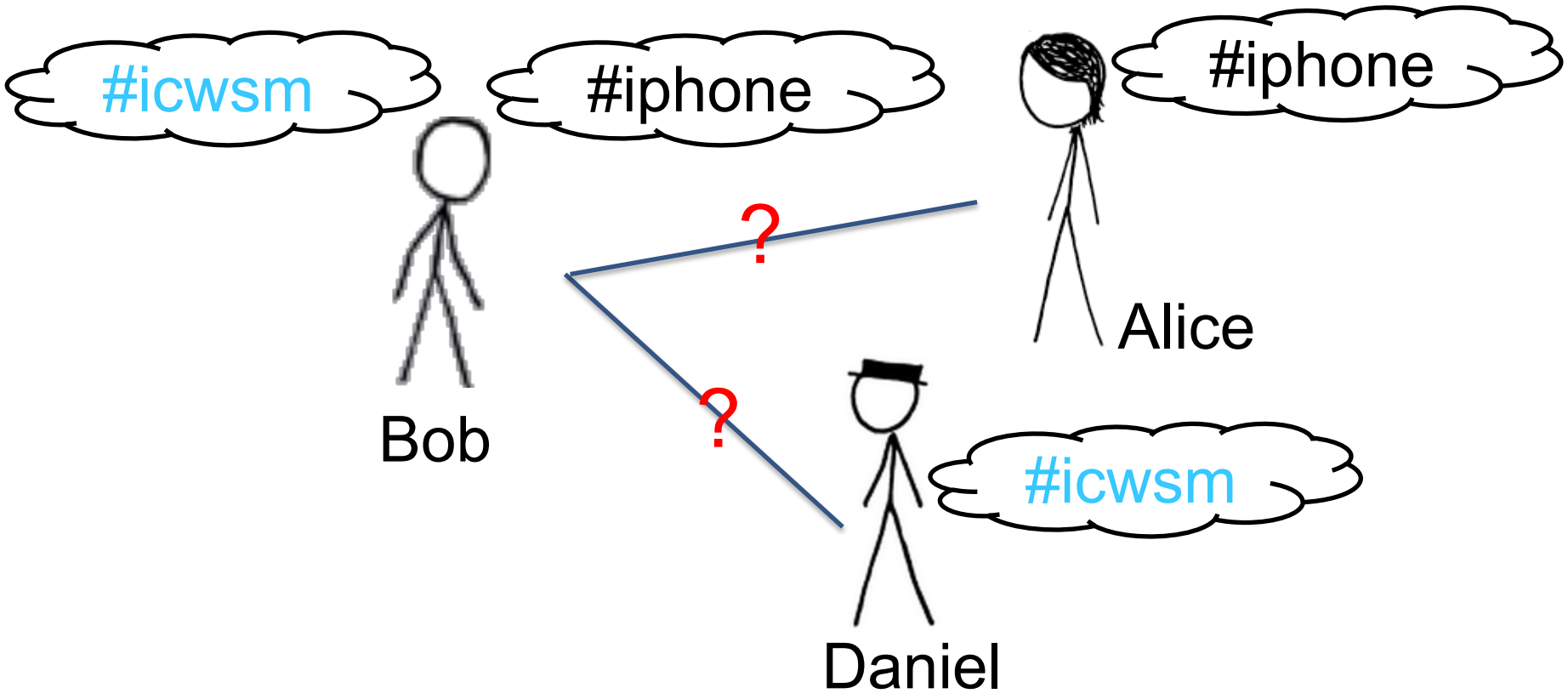
Your social relationships and your topics of interests are intuitively connected

People form friendships through mutual interests



Your social relationships and your topics of interests are intuitively connected

Different topics have different predictive power about social relationships



Research Questions

- How well can people's topics of interests predict their social relationships? [Liben-Nowell and Kleinberg 2007; Taskar et al. 2003; Schi- fanella et al. 2010; Leroy, Cambazoglu, and Bonchi 2010; Rossetti, Berlingerio, and Giannotti 2011; Hutto, Yardi, and Gilbert 2013]
- How well can the social relationships among the people interested in a topic predict the future popularity of a topic? [Lin et al. 2013]

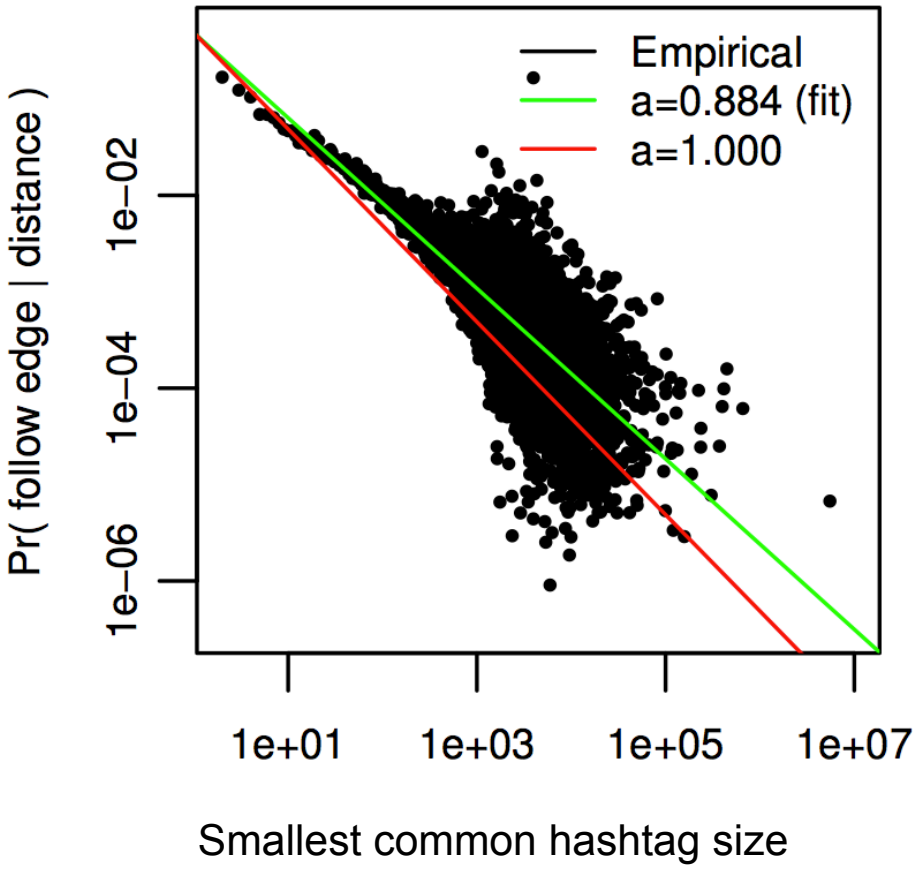
Dataset

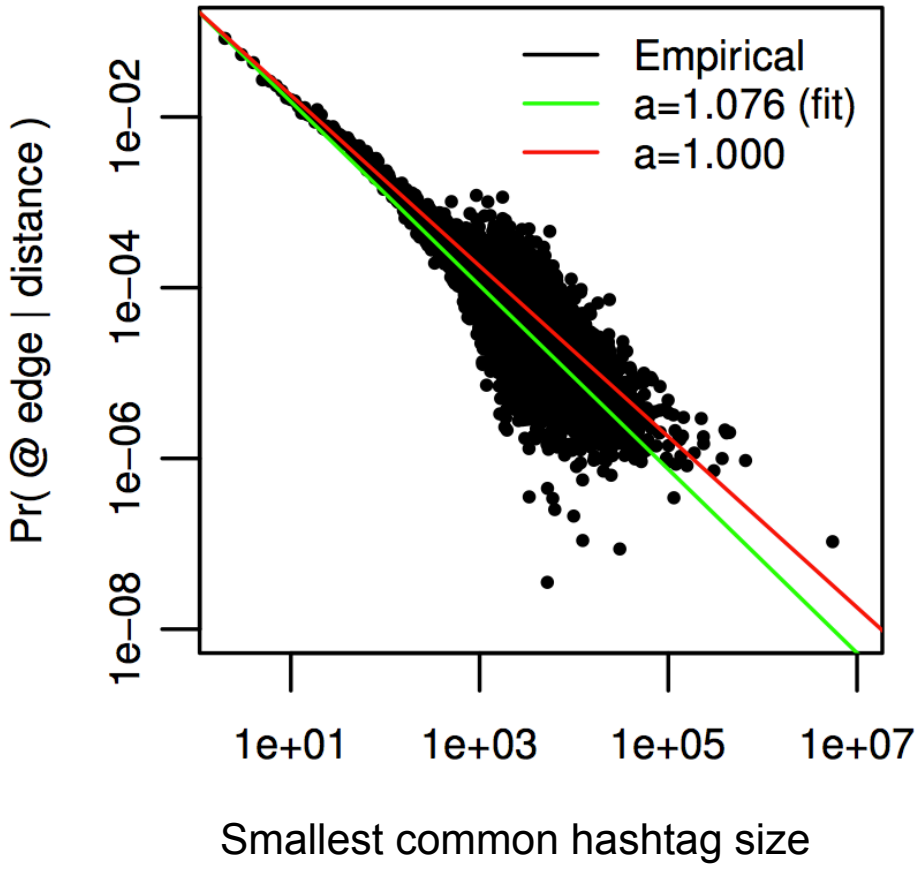
- Overview of the dataset
  - 5,513,587 users on Twitter [Romero, Meeder, and Kleinberg 2011]
  - 7,305,414 unique hashtags (topics)
  - Graphs
    - Follow graph: 366M follow edges [Kwak et al. 2010]
    - @ graph: 85M @-edges

    A has an @-edge to B, if A @-mentions B in at least 1 tweet (threshold=1, we will try different thresholds in later experiments)

Link probability vs Smallest common hashtag size (log-log)

Hashtag size: the number of users who have used a certain hashtag



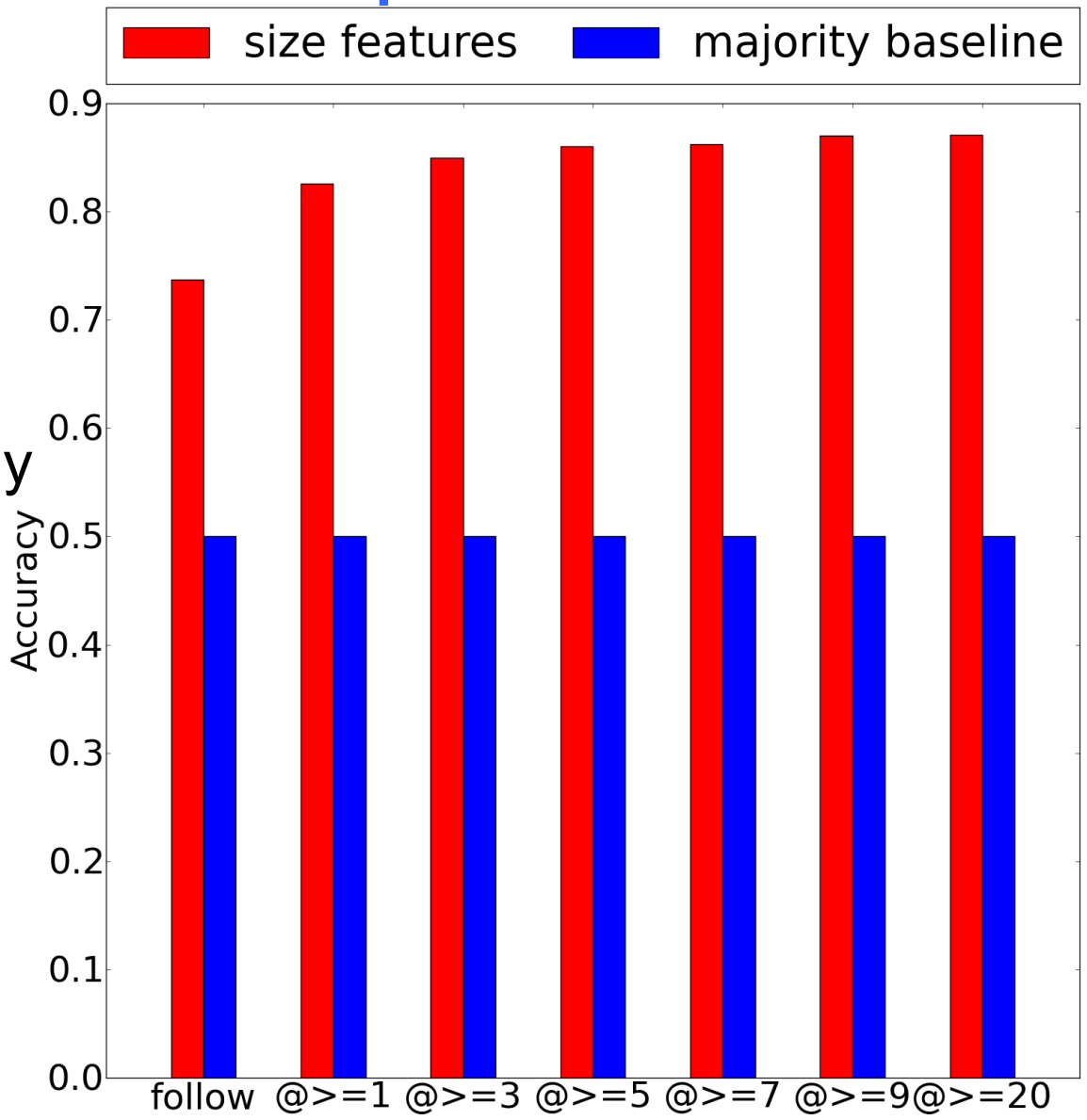


Predicting social relationships

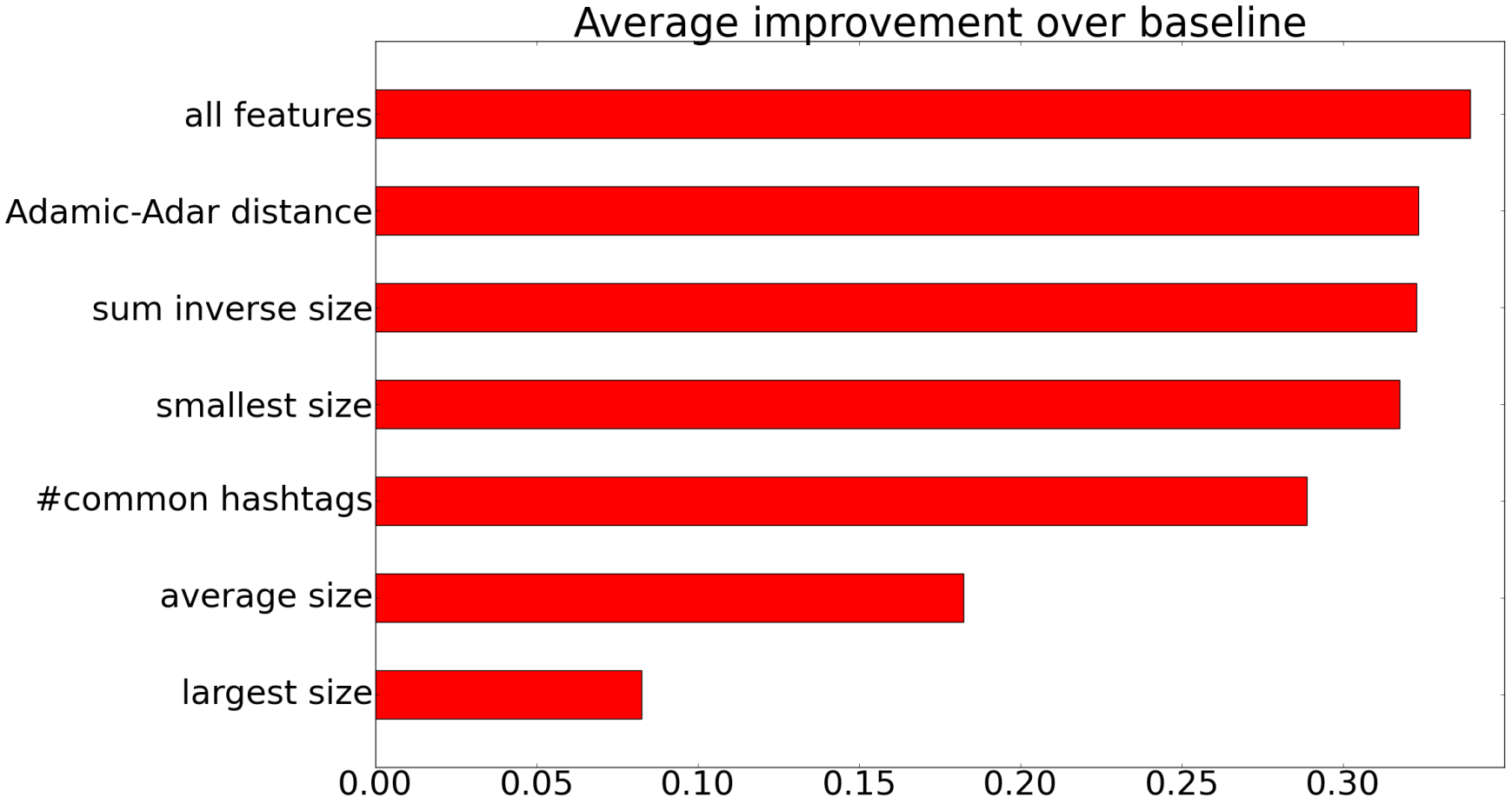
- Predict the presence of edges
- Balanced prediction task
  - 50,000 connected pairs, 50,000 disconnected pairs
- Features based on hashtag sizes
  - number of hashtags in common
  - size of the smallest common hashtag
  - size of the largest common hashtag
  - average size of the common hashtags
  - sum of the inverse sizes ($\Sigma_h 1/|h|$)
  - Adamic-Adar distance, adapted to hashtags ($\Sigma_h 1/\log(|h|)$)
- Logistic regression, 10-fold cross validation

Performance on Predicting Social Relationships

- Using basic hashtag size features can predict social relationships accurately
- Strong ties are easier to predict



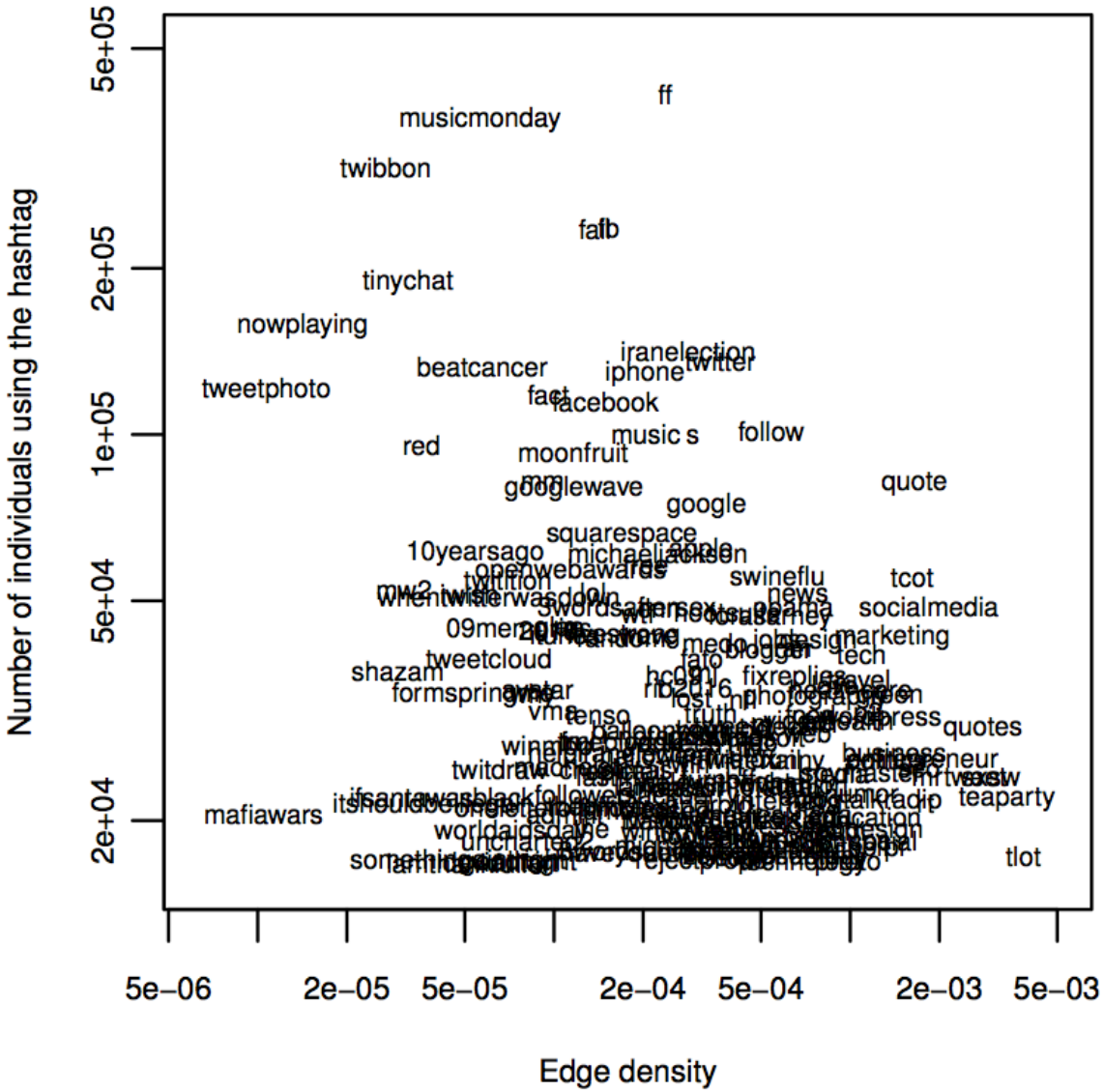
Performance of a Single Feature



- Adamic-Adar distance and sum of inverse sizes are the best single features
- Smallest common hashtag size is quite good as such an simple feature

Beyond Hashtag Size

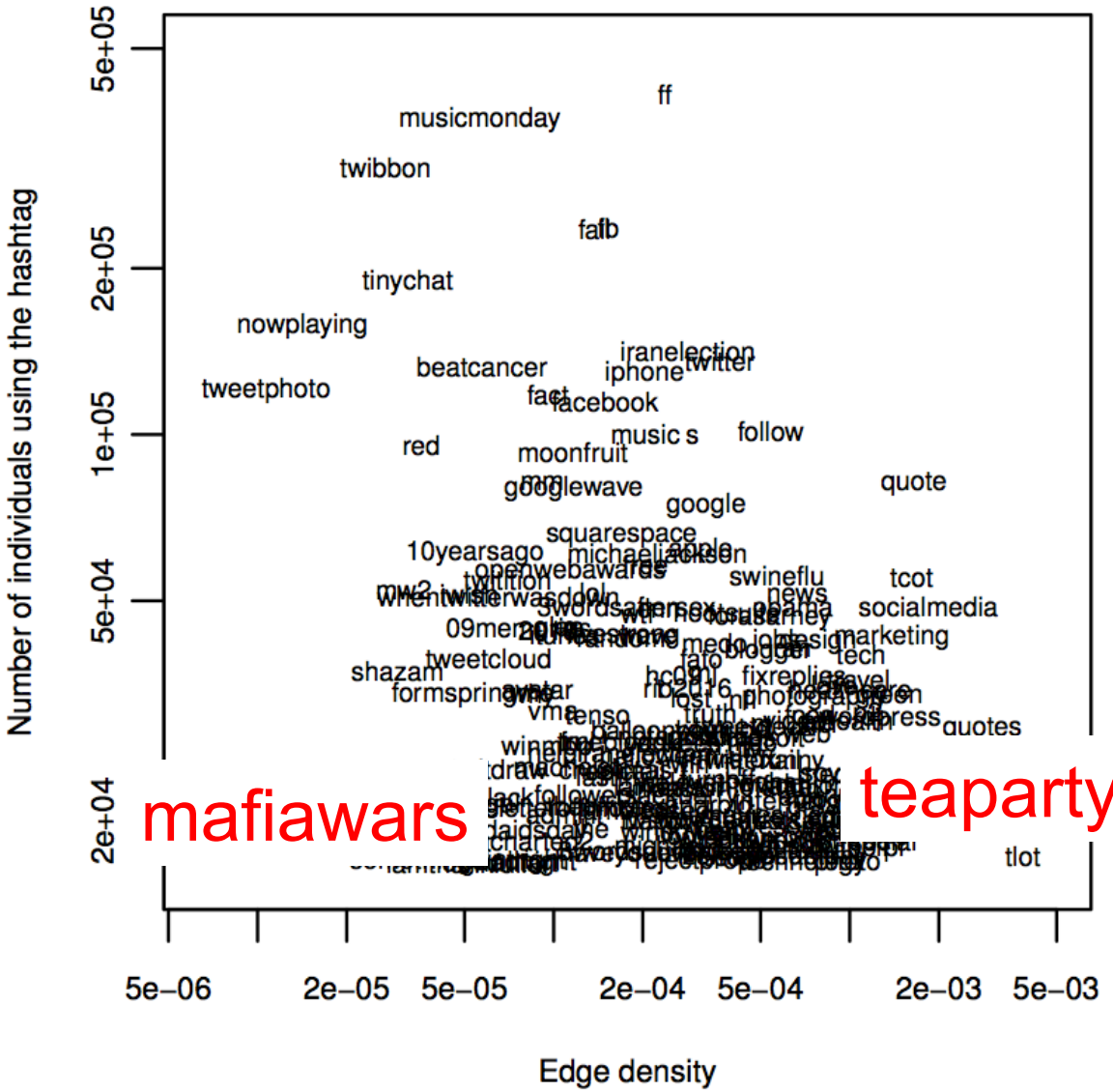
Edge density heterogeneity for the 200 most popular hashtags (edge density=|E|/(|V|*(|V|-1)))



Beyond Hashtag Size

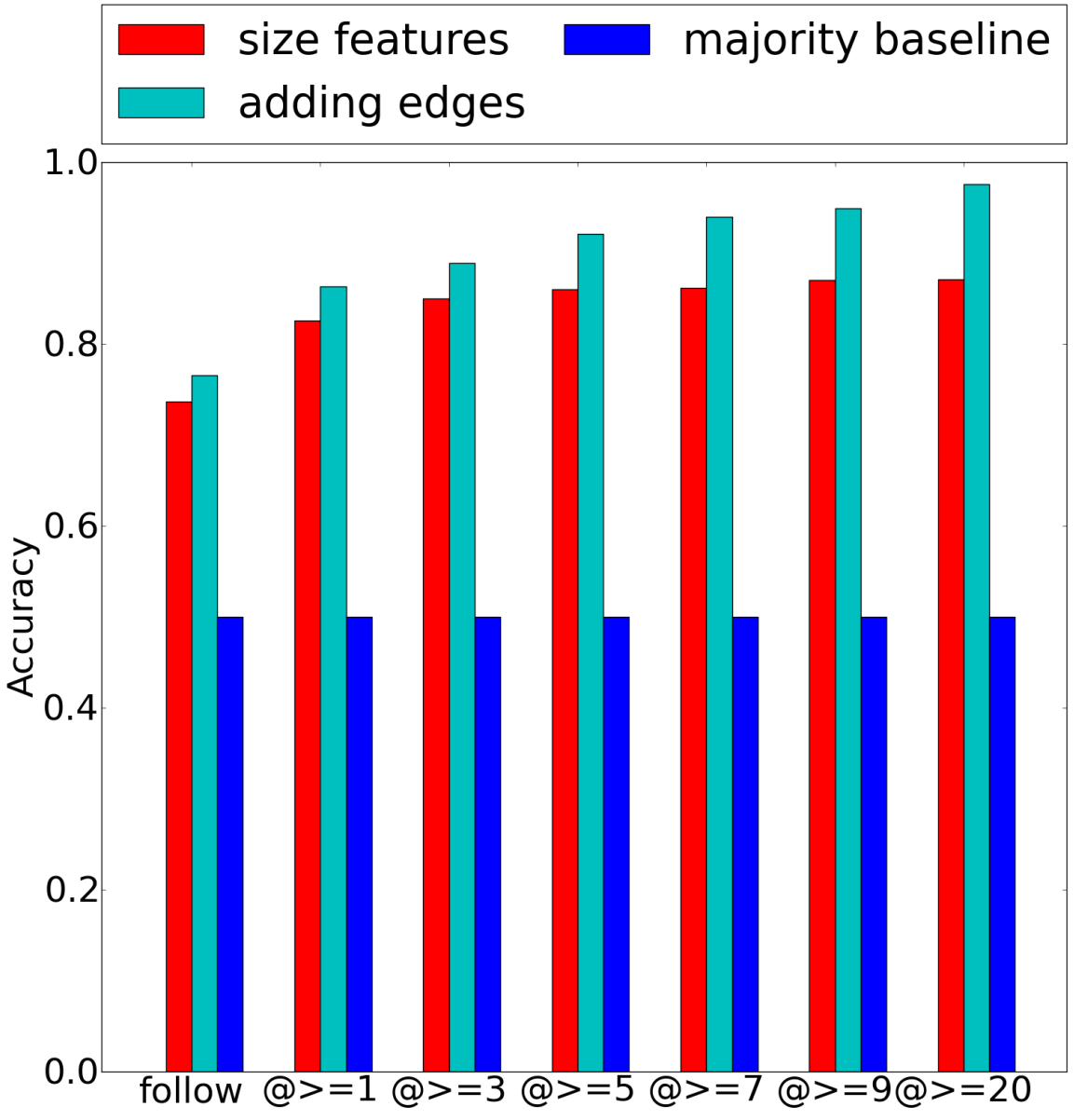
For hashtags with the same size, the connections between adopters are quite different
e.g., #mafiawars vs #teaparty

Add feature: Number of edges between the users *other than the user pair that is being predicted* who used the smallest common hashtag



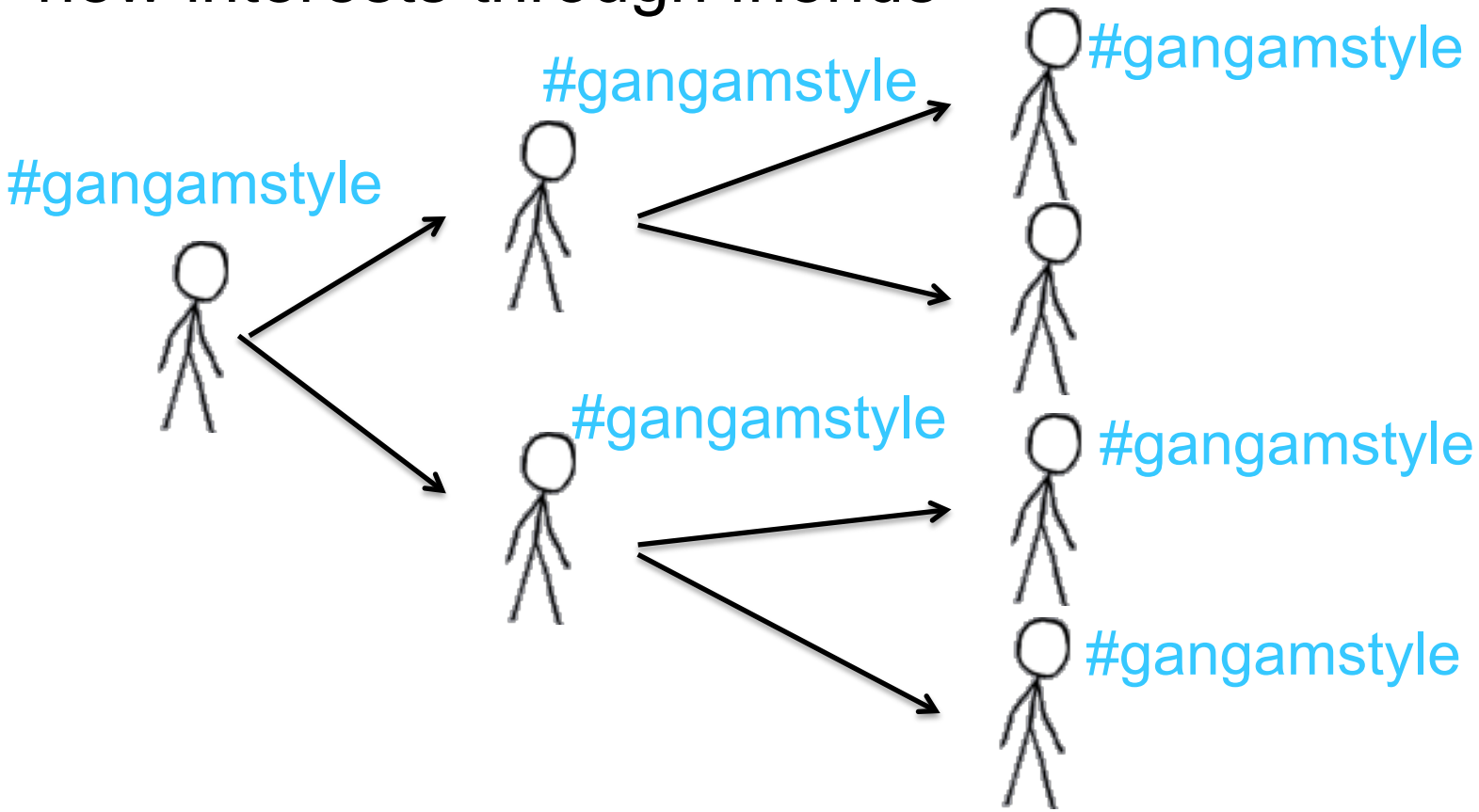
Adding Graph Information

- The best performance is achieved with adding graph information
- The improvement is much larger for strong ties



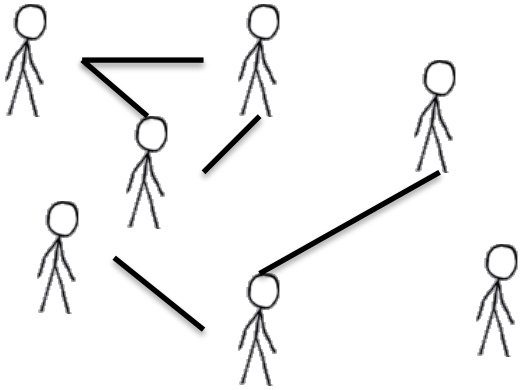
Part II: From Social Structure to Topical Structure

Word of mouth: People can discover new interests through friends



How well can the social relationships among the people interested in a topic predict the future popularity of a topic?

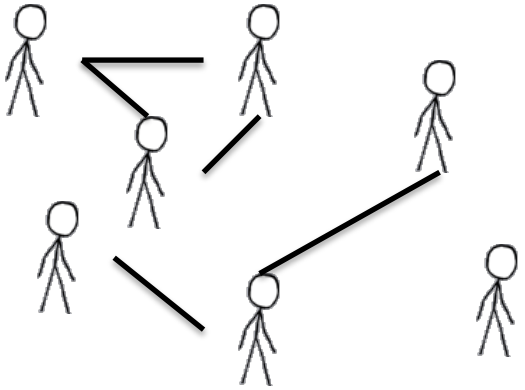
Graph structure of the initial adopters of #gangamstyle



Future popularity of #gangamstyle

How well can the social relationships among the people interested in a topic predict the future popularity of a topic?

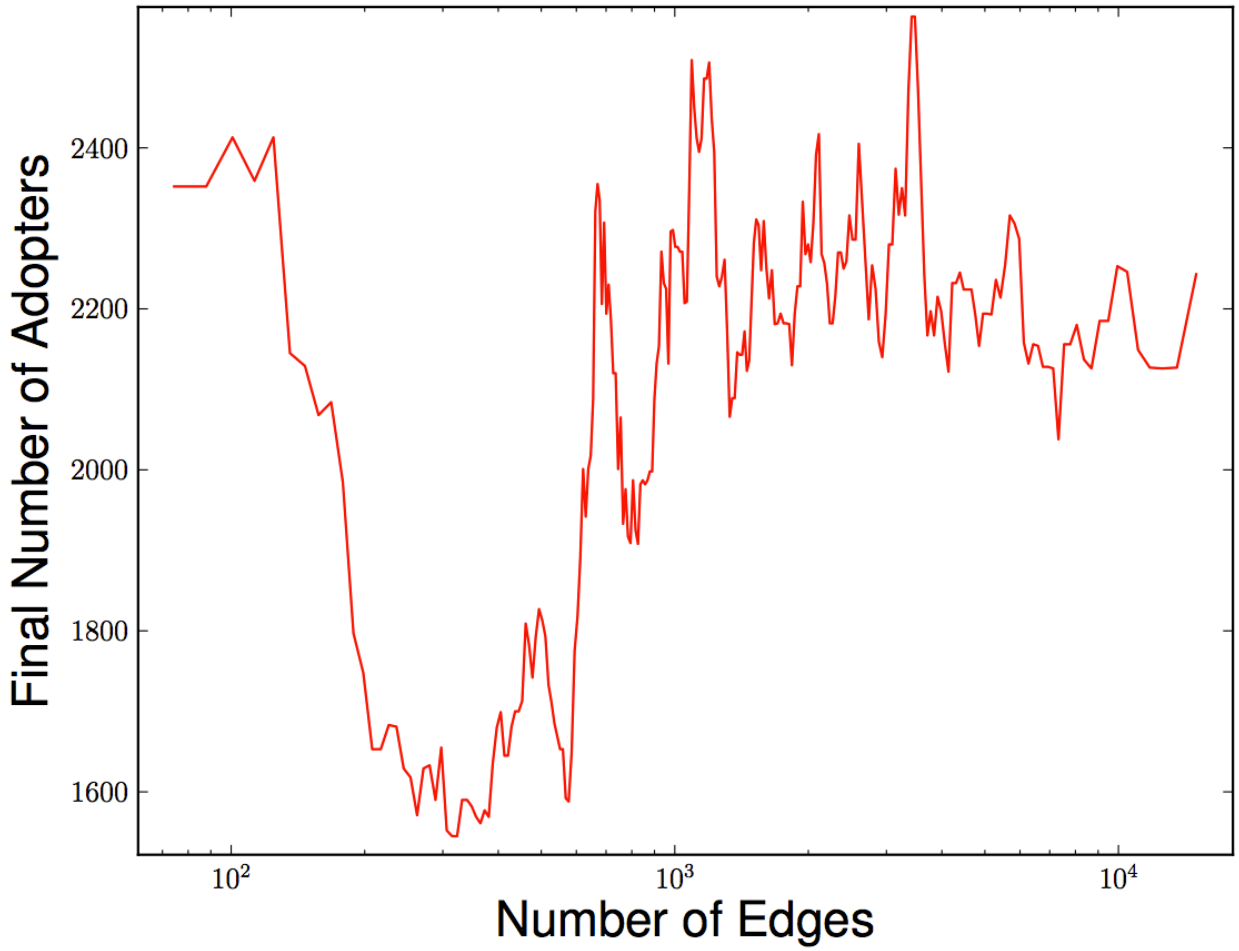
Graph structure of the initial adopters of #gangamstyle



Future popularity of #gangamstyle

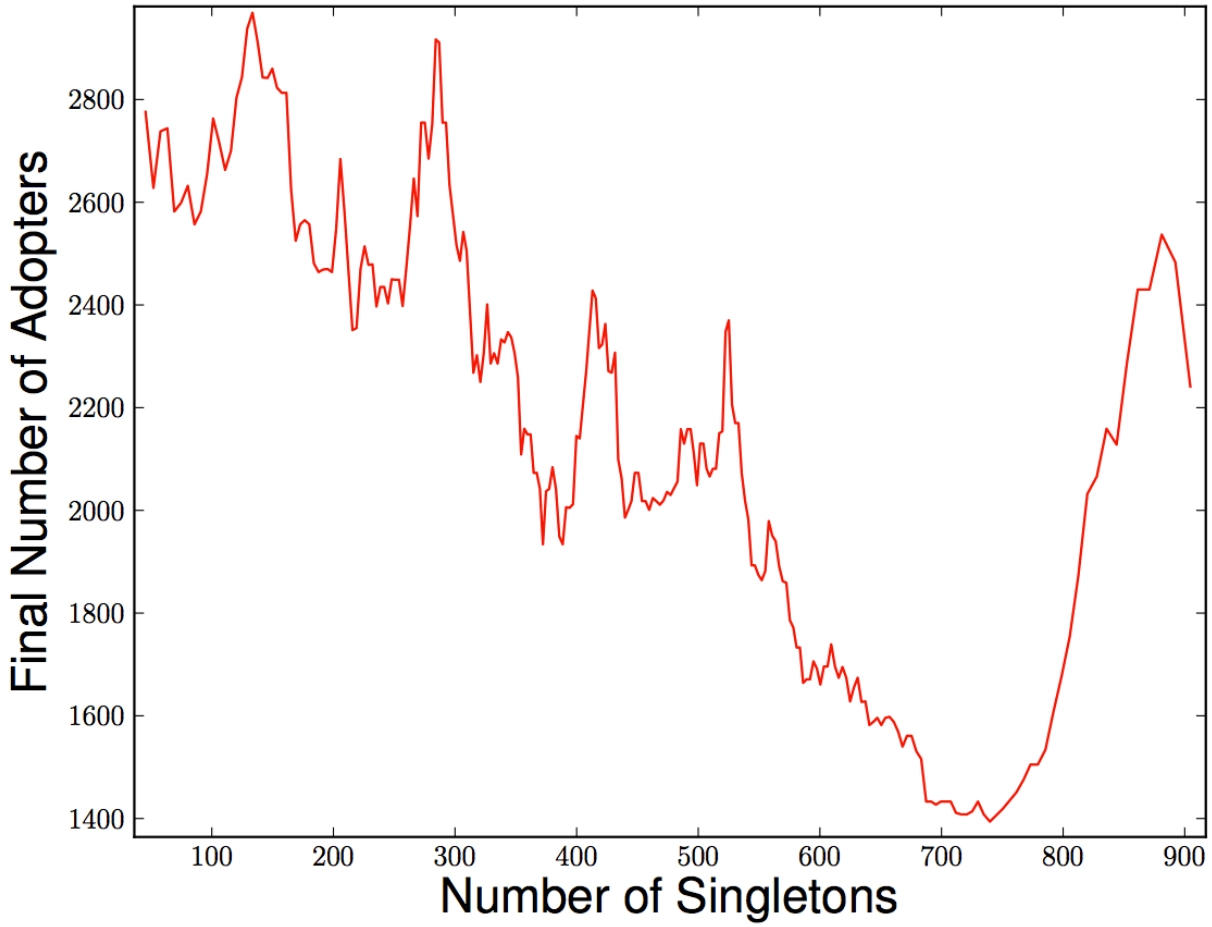
Data: 7,397 hashtags that had at least 1,000 adopters

Eventual popularity vs number of edges in the first 1000 adopters



It is not monotone, there is an interior minimum

Eventual popularity vs number of singletons in the first 1000 adopters



Again, an interior minimum on the right!

High-level Intuitions on Interior Minimum

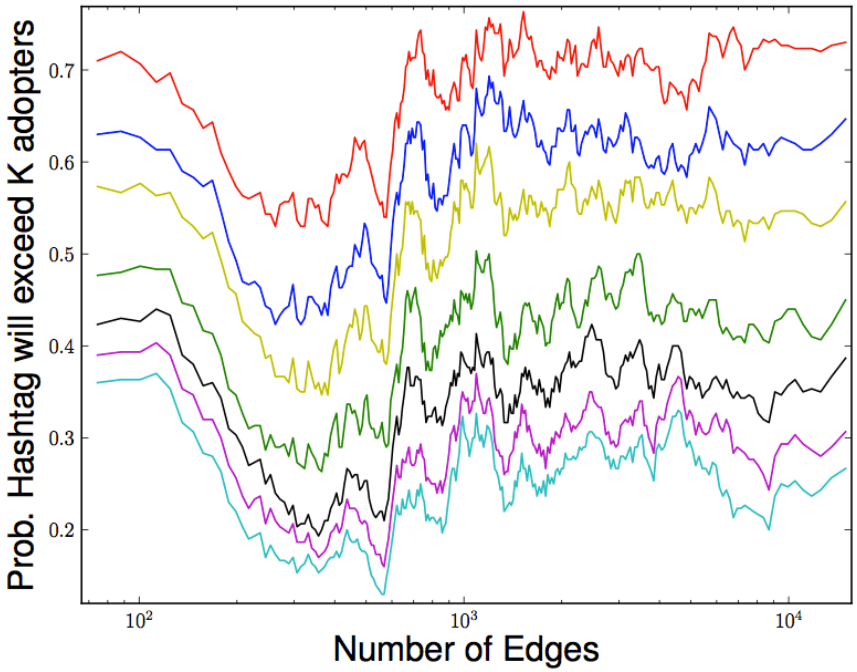
- If the initial adopters are very well connected, the topics have a better chance to be viral

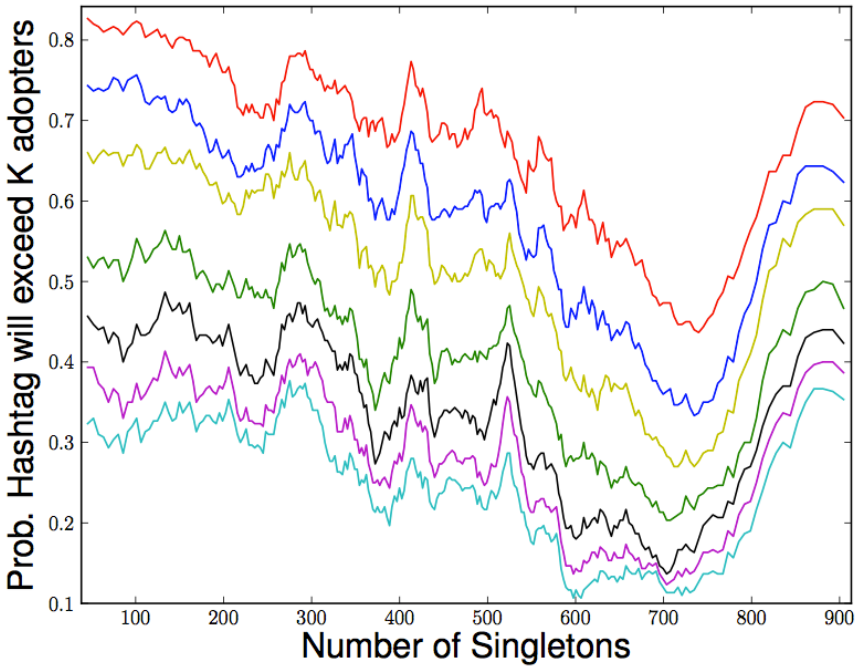
  e.g., #tcot, #tlot
- If the initial adopters are totally disconnected, the topics are probably related to exogenous events, and they can become popular

  e.g., #iphone, #michaeljackson, #bigbird

Probability that hashtag size will exceed K users

K = 1500, 1750, 2000, 2500, 3000, 3500, 4000





- The trend is consistent no matter what K is
- There is an interior minimum

Prediction Task

- Predict whether the eventual size will double (K->2K)
- Using features from the subgraph induced by the first K adopters (follow vs @>=3)

Features of Subgraphs

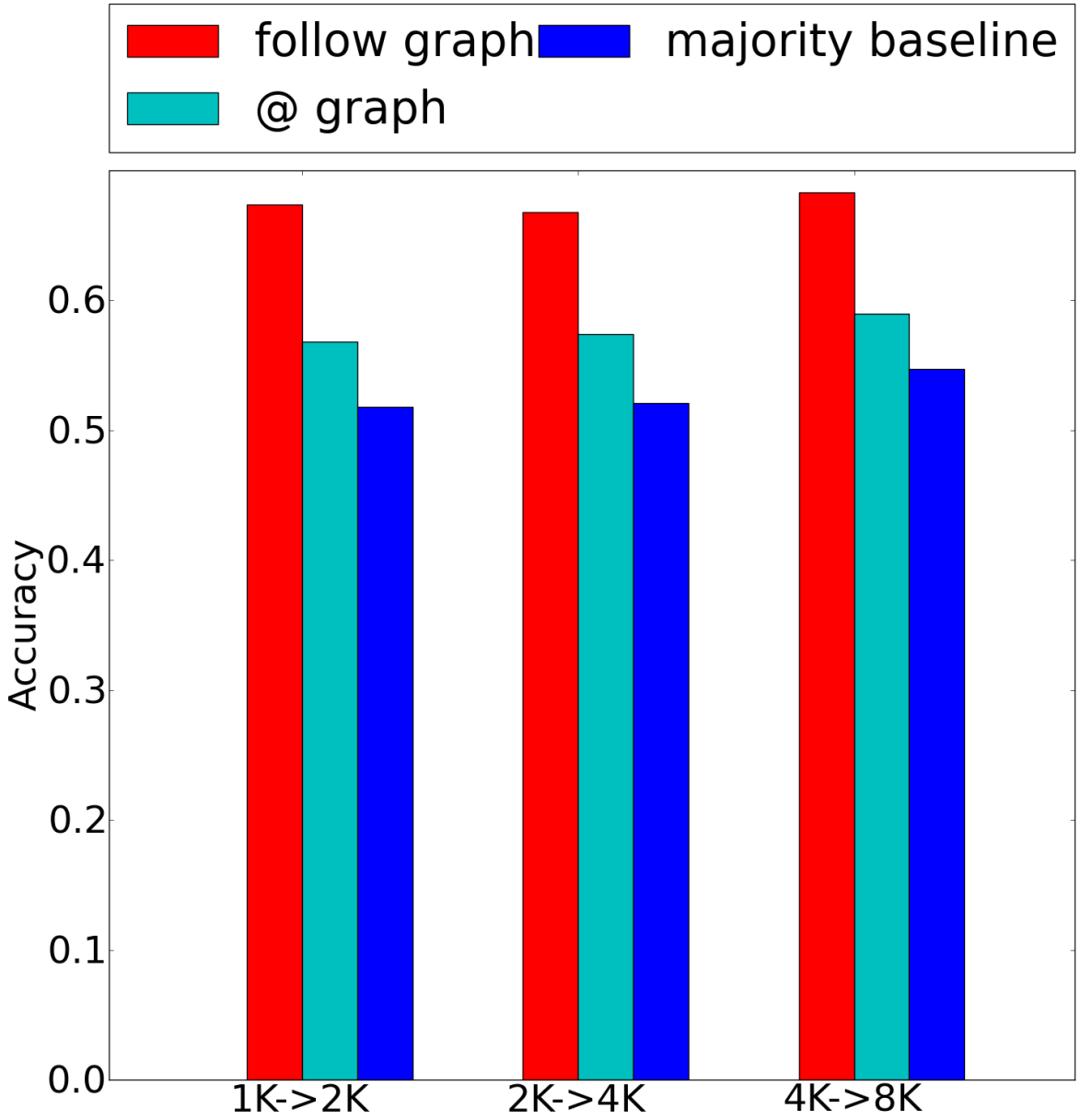
- Number of edges
- Number of singletons
- Number of (weakly) connected components
- Size of the largest (weakly) connected component
- Raw value, log(value)

Features of Subgraphs

- Number of edges
- Number of singletons
- Number of (weakly) connected components
- Size of the largest (weakly) connected component
- Raw value, log(value), |*value-(max value / 2)*|

Performance on Predicting Popularity

- The performance with graph features is much better than majority baseline
- Using follow graph is better than @ graph



Summary

- Merely basic features from topical structures can predict social relationships accurately
- The connections between early adopters can predict the eventual popularity of the topic
- Strong ties are the easiest to predict from hashtag structure, but they are much less useful in predicting the hashtag popularity

Summary

- Merely basic features from topical structures can predict social relationships accurately
- The connections between early adopters can predict the eventual popularity of the topic
- Strong ties are the easiest to predict from hashtag structure, but they are much less useful in predicting the hashtag popularity

Thank you! & Questions?

Chenhao Tan
chenhao@cs.cornell.edu
@ChenhaoTan