Online Social Network Research: A Case Study of CyWorld

CUHK/ASTRI visit

2009.7.16. Thursday

Sue Moon

文壽福

Stanley Milgram's Small-World Experiment

- In 1967 he sent packages to 160 people in Omaha, Nebraska
- He asked them to forward towards a stockbroker in Boston

$\Rightarrow$ Those packages that came back had 6 degrees of separation

Leskovec and Horvitz' MSN analysis

- MSN network of June 2006
- 180 of 240 million accounts active
- 1.3 billion edges

$\Rightarrow$ Avg shortest distance = 6.6 (median 6)

Blossoming Social Technologies

Motivations

- Emergence of online social networking services drew much attention in CS
- OSN poses challenges/opportunities in
  - malware epidemics
  - information diffusion
  - recommendation engine design
  - social network evolution
  - distributed system/database design and mgmt

Overview of Today's Talk

- Analysis of CyWorld Ilchon Network (WWW'07)
- Comparison of CyWorld Ilchon and Guestbook Networks (IMC'08)
- Consistent community identification (IMC'09)

Past and Ongoing Collaborators

- Physics (Prof. Hawoong Jeong)
  - Yong-Yeol Ahn (PhD; now post-doc at Northeastern)
  - Young-Ho Eom (PhD; now post-doc at KAIST)
- Computer Science
  - Seungyeop Han (MS; now at Naver)
  - Hyunwoo Chun (PhD)
  - Yoonchan Choi (MS; now at Samsung)
  - Haewook Kwak (PhD)

Part I:
Analysis of
CyWorld Ilchon Network

CyWorld

- Largest SNS in South Korea
  - Started in September 2001
  - 10 million users in 2004
  - Over 20 million users out of 49 million population
- Front runner of many features
  - Friend (*ilchon*) relationship
  - Guestbook (Facebook Wall)
  - Testimonial (*il-chon-pyung*)
  - Photos - scraps
  - Avatar in cyber home

My CyWorld “Mini-Homepage”

Overview of the Talk

- Metrics representative of topological characteristics
- Analysis of CyWorld, MySpace, orkut
- Future work

CyWorld Data Set

- Complete snapshot (Nov 2005)
  - 191 million friend relationships between 12 million users

Metrics of Interest

- Degree distribution
  - "Power-law"
    - Small number of nodes have large numbers of links
- Clustering coefficient *C(k)*
  - # of existing links / # of all possible links between a link's adjacent neighbors
    - Close to 1, close to a mesh
- Degree correlation k_{nn}
  - Degree k ~ mean degree of adjacent neighbors of nodes with degree k
  - Assortativity: characteristic of k_{nn} distribution

Questions We Raise

- What are the main characteristics of online SNSs?
- How representative is a sampled network?
- How does a social network evolve?

Degree Distribution

Two scaling regions

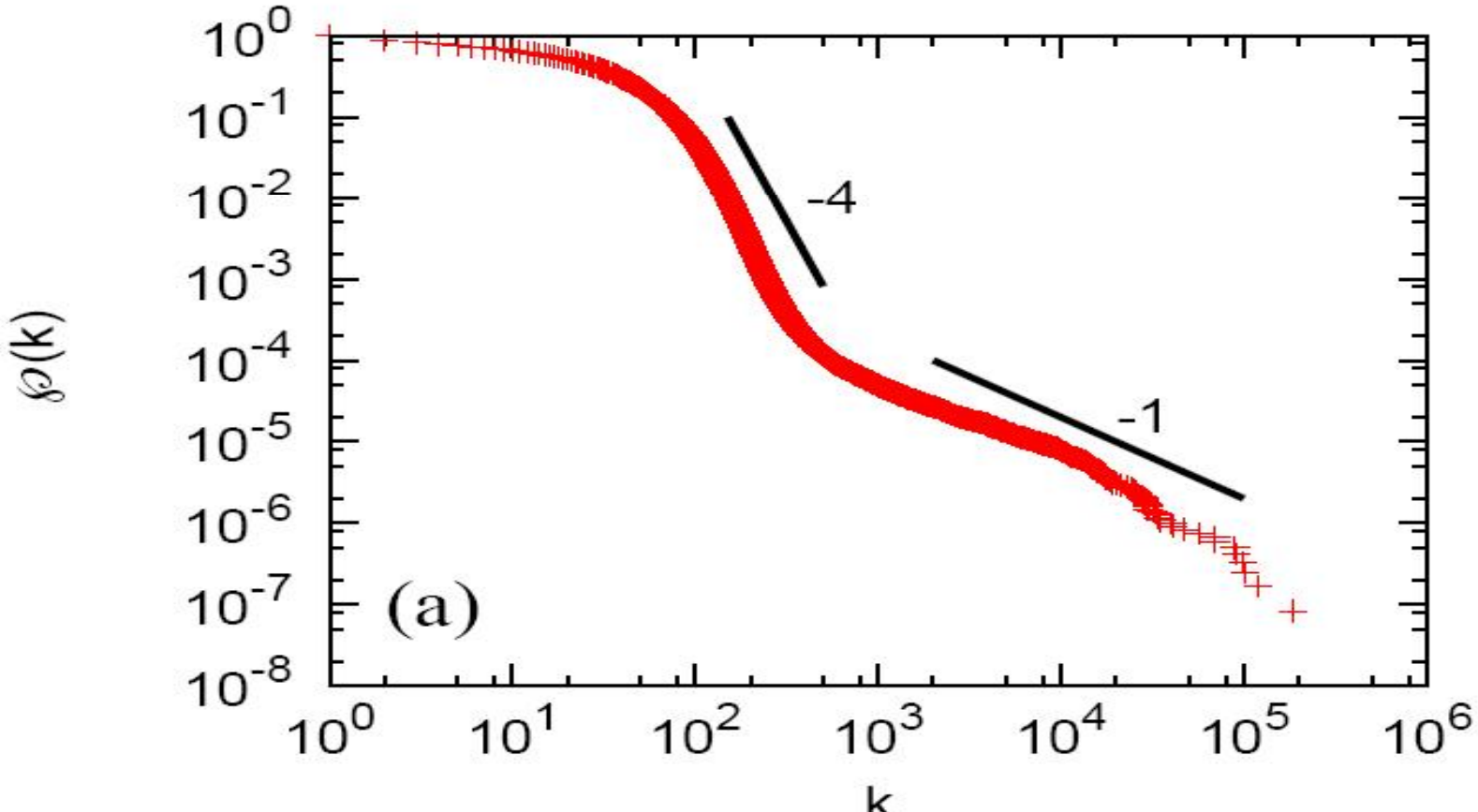
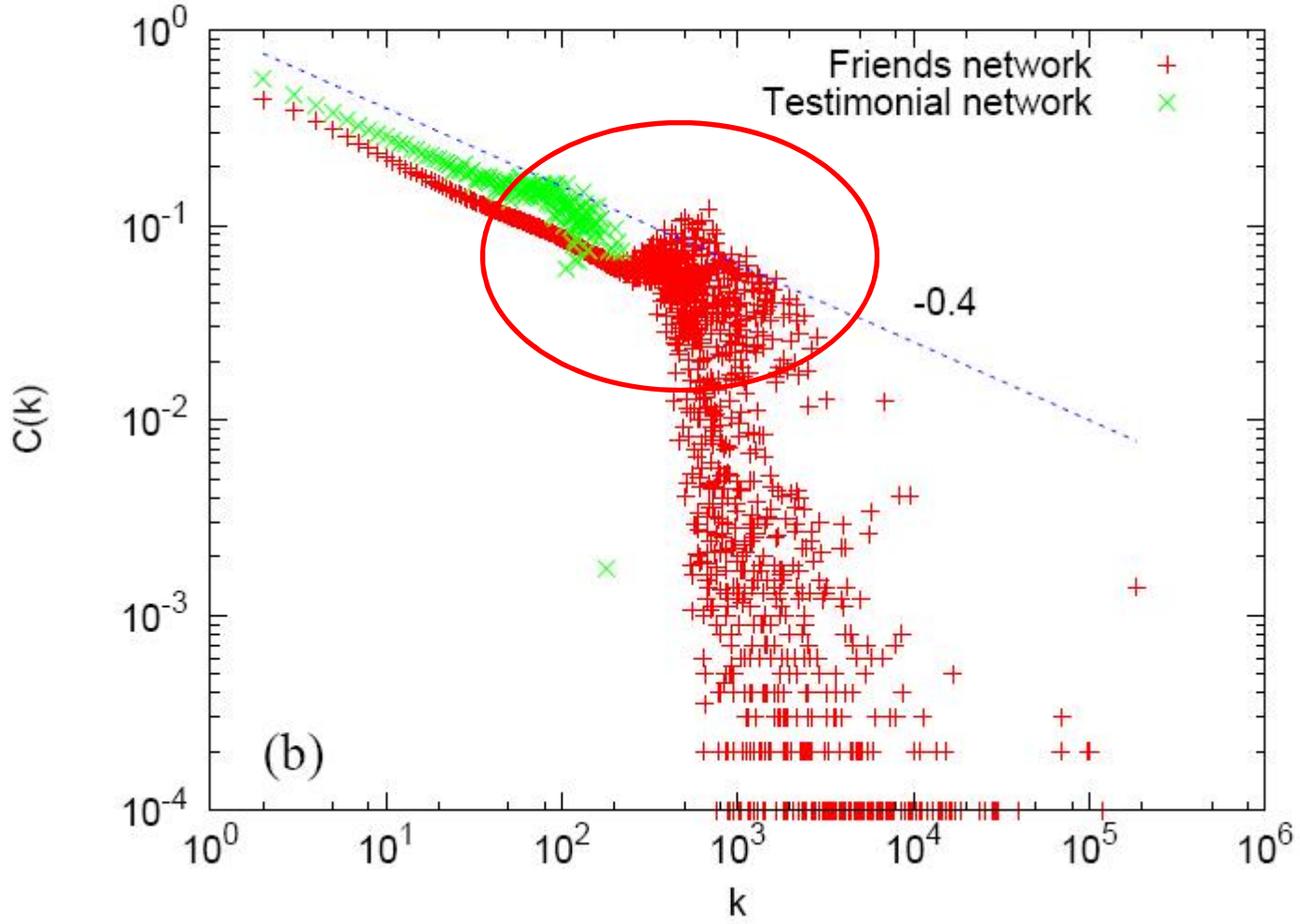
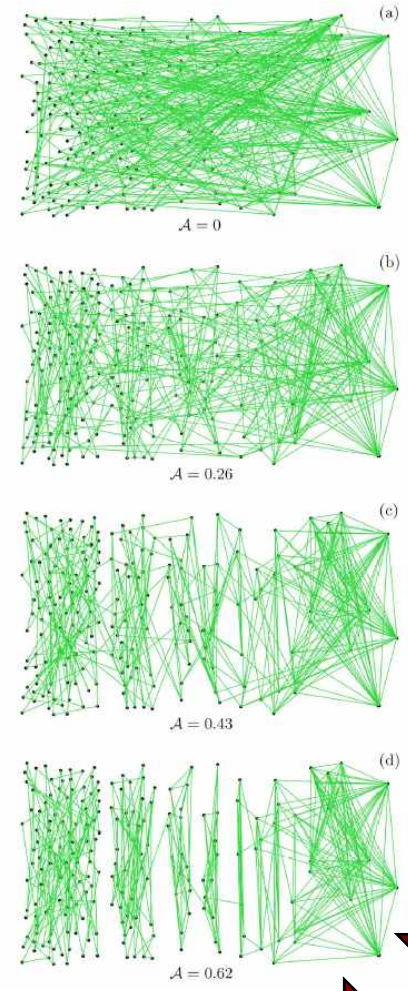


Figure 1-(a): degree distribution, CCDF

Clustering Coefficient Distribution



Assortative Mixing



assortative

degree

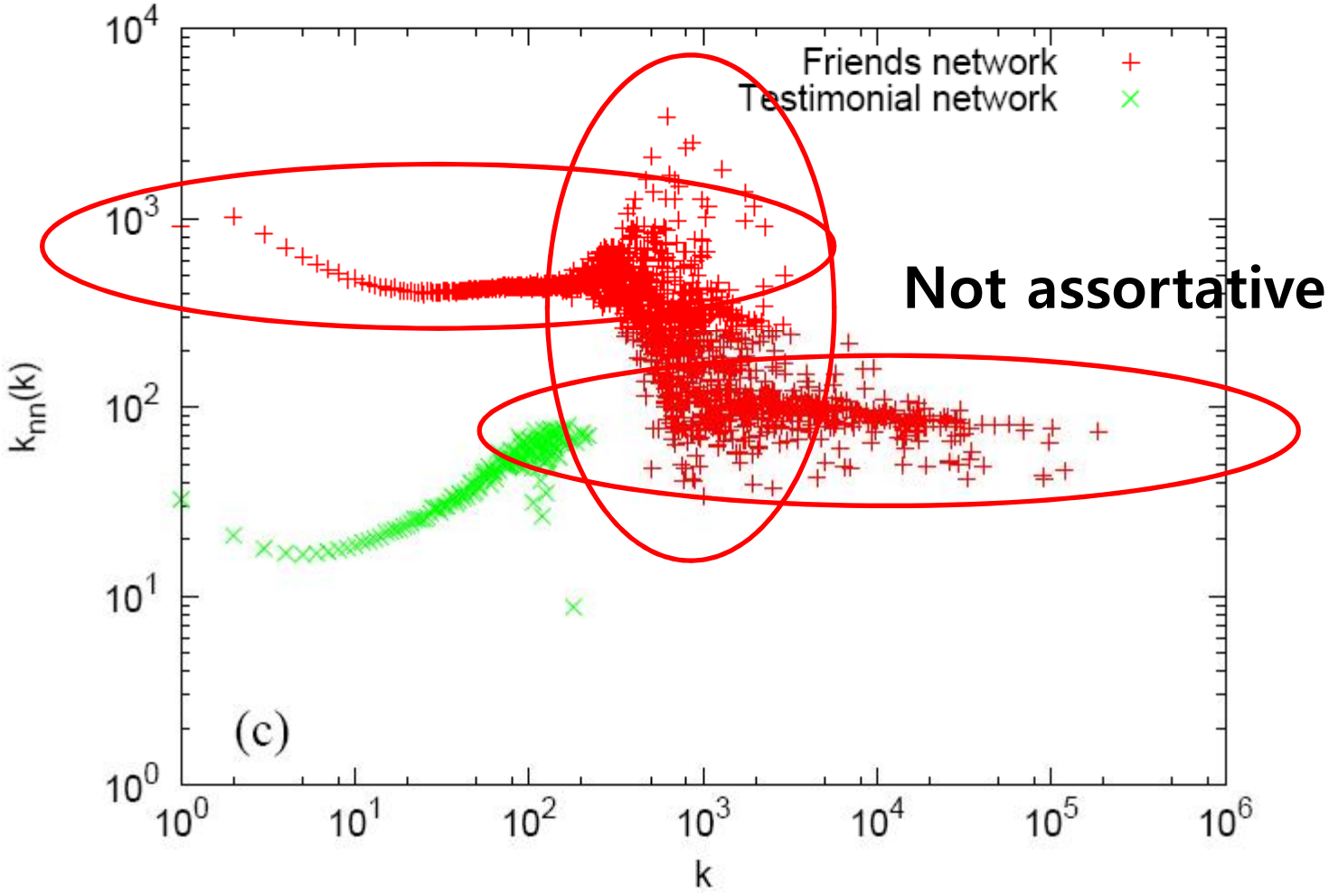
"Social"

"non-social"

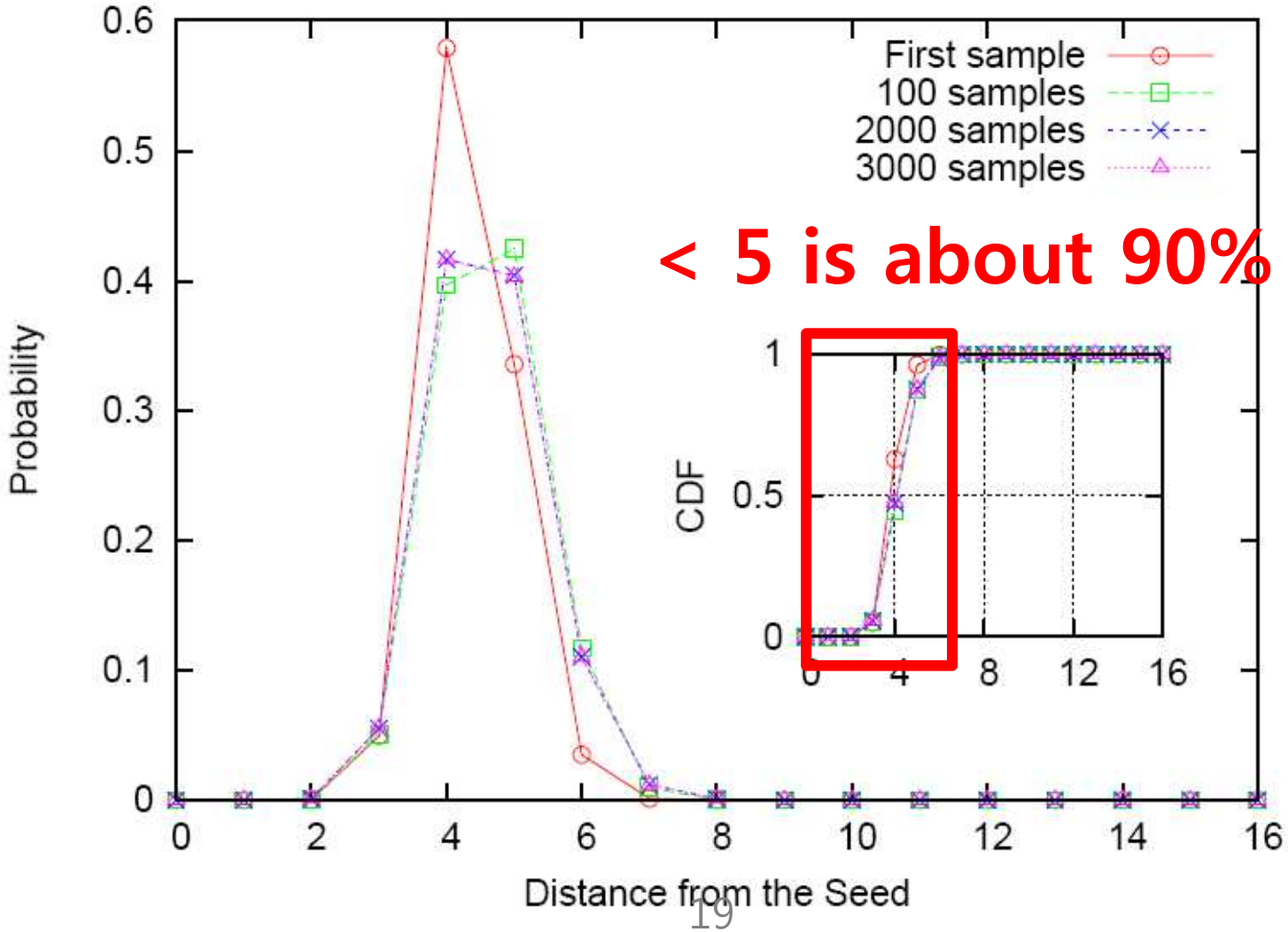
Network	n	r	
Physics coauthorship (a)	52 909	0.363	+
Biology coauthorship (a)	1 520 251	0.127	+
Mathematics coauthorship (b)	253 339	0.120	+
Film actor collaborations (c)	449 913	0.208	+
Company directors (d)	7 673	0.276	+
Internet (e)	10 697	−0.189	−
World-Wide Web (f)	269 504	−0.065	−
Protein interactions (g)	2 115	−0.156	−
Neural network (h)	307	−0.163	−
Marine food web (i)	134	−0.247	−
Freshwater food web (j)	92	−0.276	−
Random graph (u)		0	
Callaway *et al.* (v)		$\delta/(1+2\delta)$	
Barabási and Albert (w)		0	

M. E. J. Newman, Phys. Rev. Lett. 89, 208701 (2002)

Degree Correlation



Average Path Length



MySpace Data Set

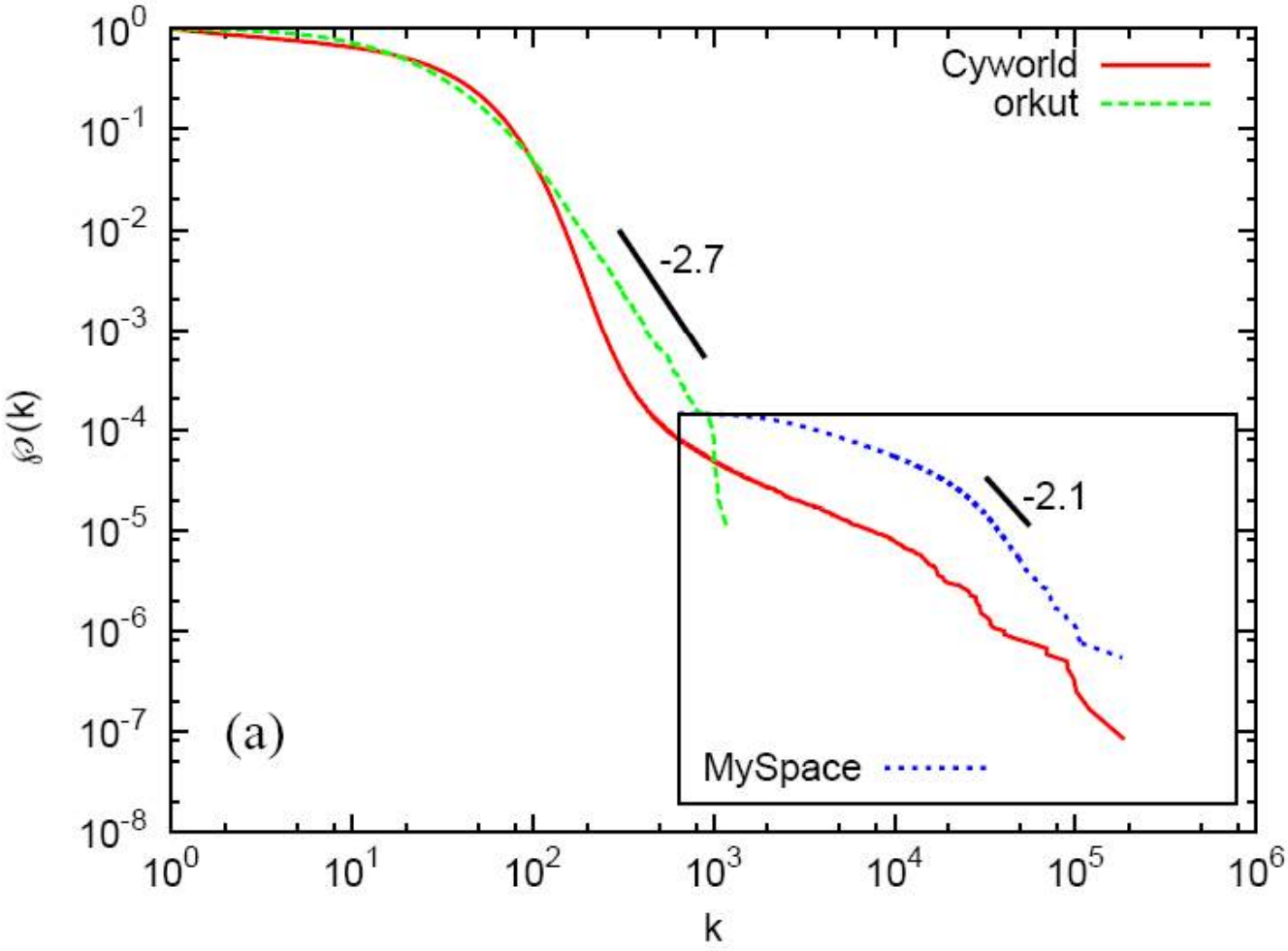
- Largest in the world
  - Began in Jul 2003
  - Has 130 million by Nov 2006
- Snowball sampled
  - During Sep/Oct 2006
  - Random seed to 100,000 users
  - About 23% of users had friend list hidden

Orkut Data Set

- Google SNS
  - Began in Sep 2002
  - Became official Google service in Jan 2004
  - Began as invitation-only; open now
  - Has 33 million users

- Snowball sampled
  - During Jun to Sep 2006
  - 100,000 users

Degree Distributions



Summary and Conclusions

- Cyworld has **two scaling regions** in the degree distribution.
  - Other measurements (C(k), degree correlation) also support the existence of two regions.
  - Boundary of two regions ~ 100s
    - Dunbar's Law: limit on human social relationships
  - Mature online social community
- orkut shows fast-decaying degree distribution and assortative mixing pattern – **similar to the small degree region of Cyworld.**
  - Close-knit, real-world-like social network structure.
- MySpace shows heavy tail and dissortative mixing pattern – **similar to the large degree region of Cyworld**.
  - popularity is a more important than human interaction

Future Work

- Friendship network topology not representative of activities
  - Identify steady core and study its evolution
- Explicit vs implicit communities
  - Clubs/towns vs cliquish behavior
- Growth model
  - Existing preferential attachment-based models do not fit
  - Forest fire model? Extensions?

Part II:
Comparison of
CyWorld Ilchon & Guestbook Networks

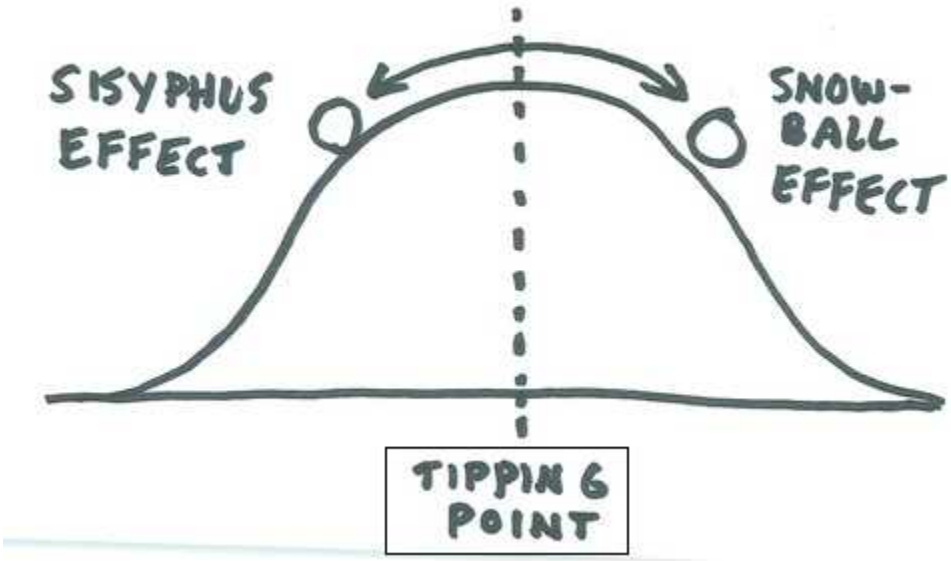
Cyworld Data Sets

- 4 snapshots of complete ilchon relationships
  - April, September, Novmber of 2005
  - November of 2006
- Guestbook logs
  - June 2003 to October 2005
  - (writer, guestbook owner, timestamp)
  - 8.4 billion messages from 17 million users

How long can Cyworld maintain exponential growth?

Tipping Point

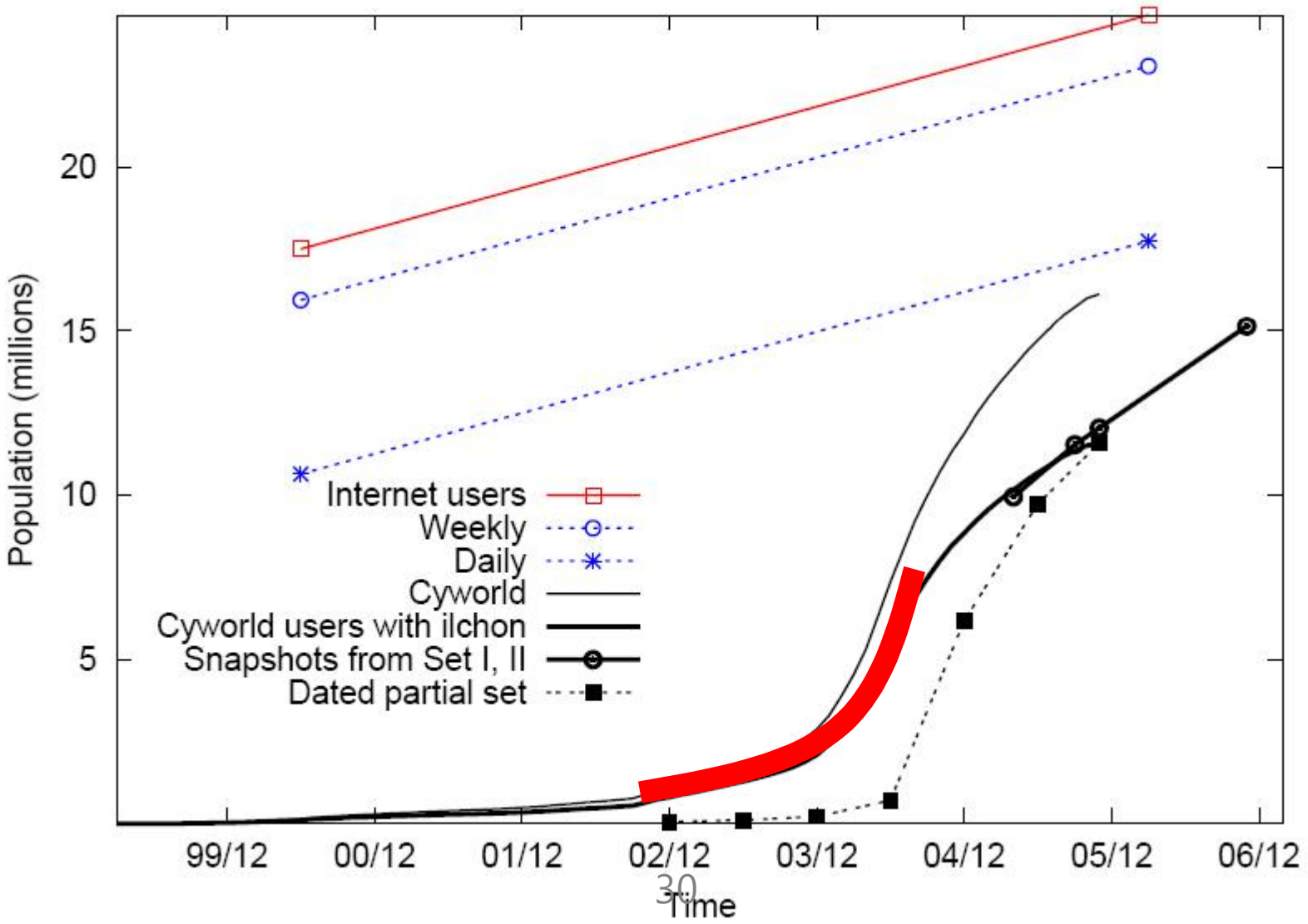
- Definition
  - Point of “phase transition” due to “peer pressure”



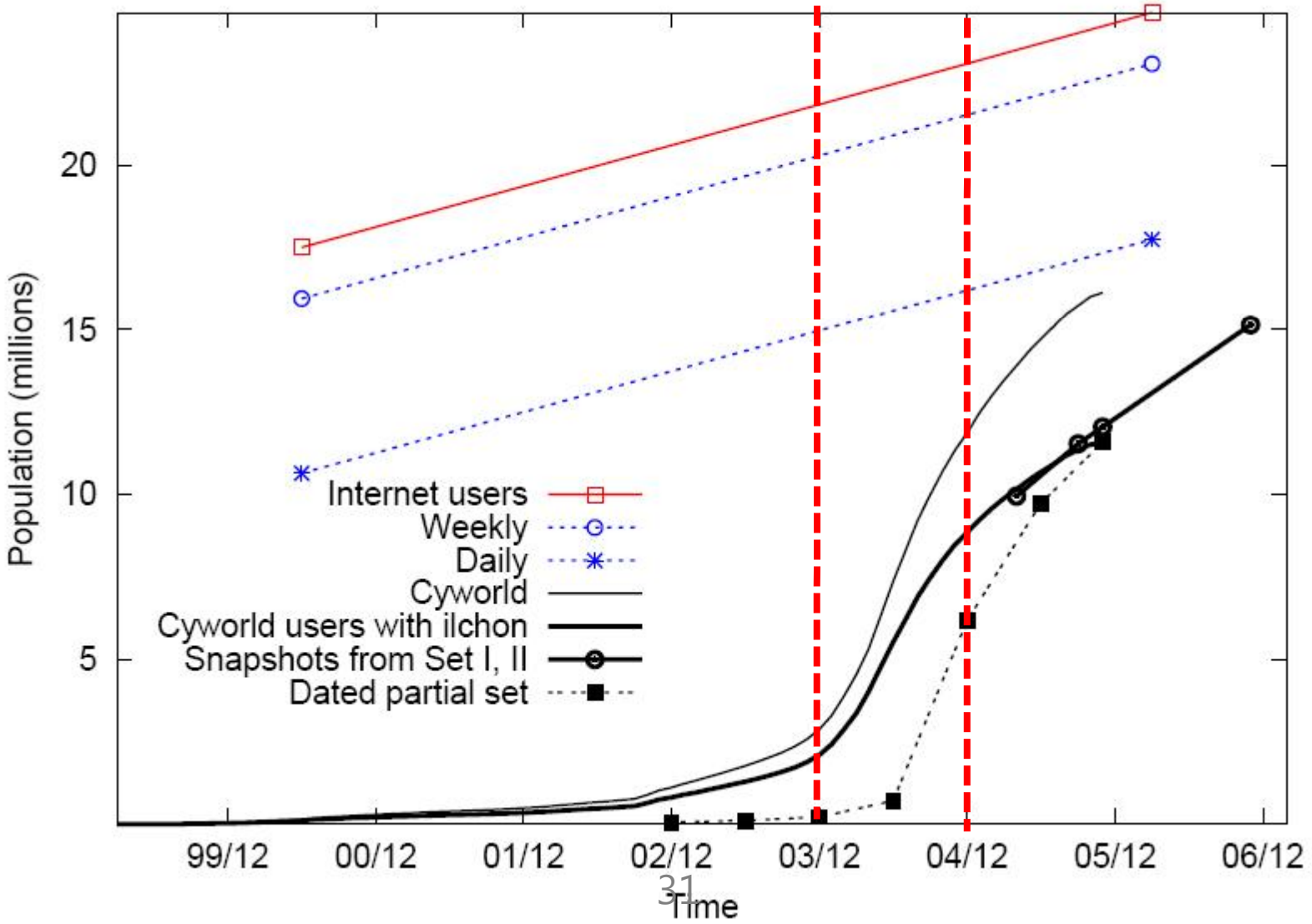
Existing Tipping Point Models

- Input parameters
  - Peer pressure
- Insufficient for our data
  - Lack of resistant factors
    - Delay in adoption
    - Individual variation in adoption
  - Still only explains our data partially

Cyworld Growth in Sheer Volume



Cyworld Growth in Sheer Volume



Shrinking Diameter [KDD2005]

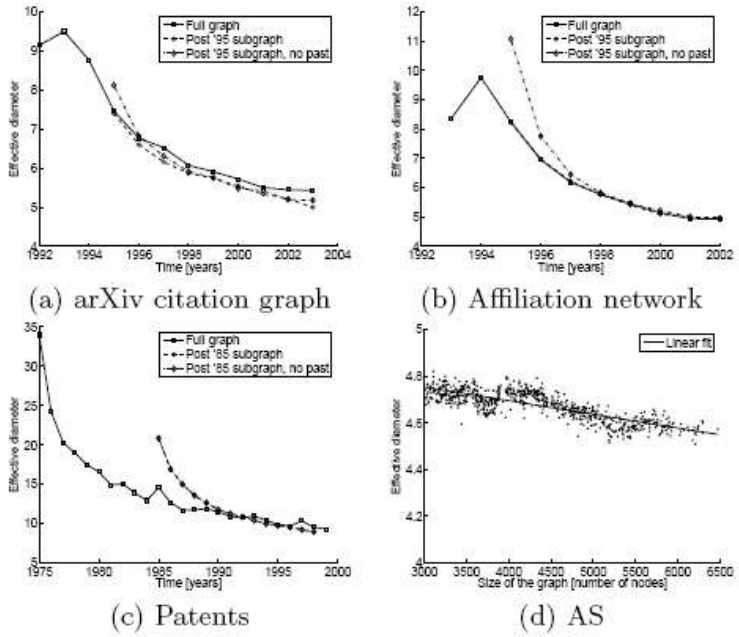
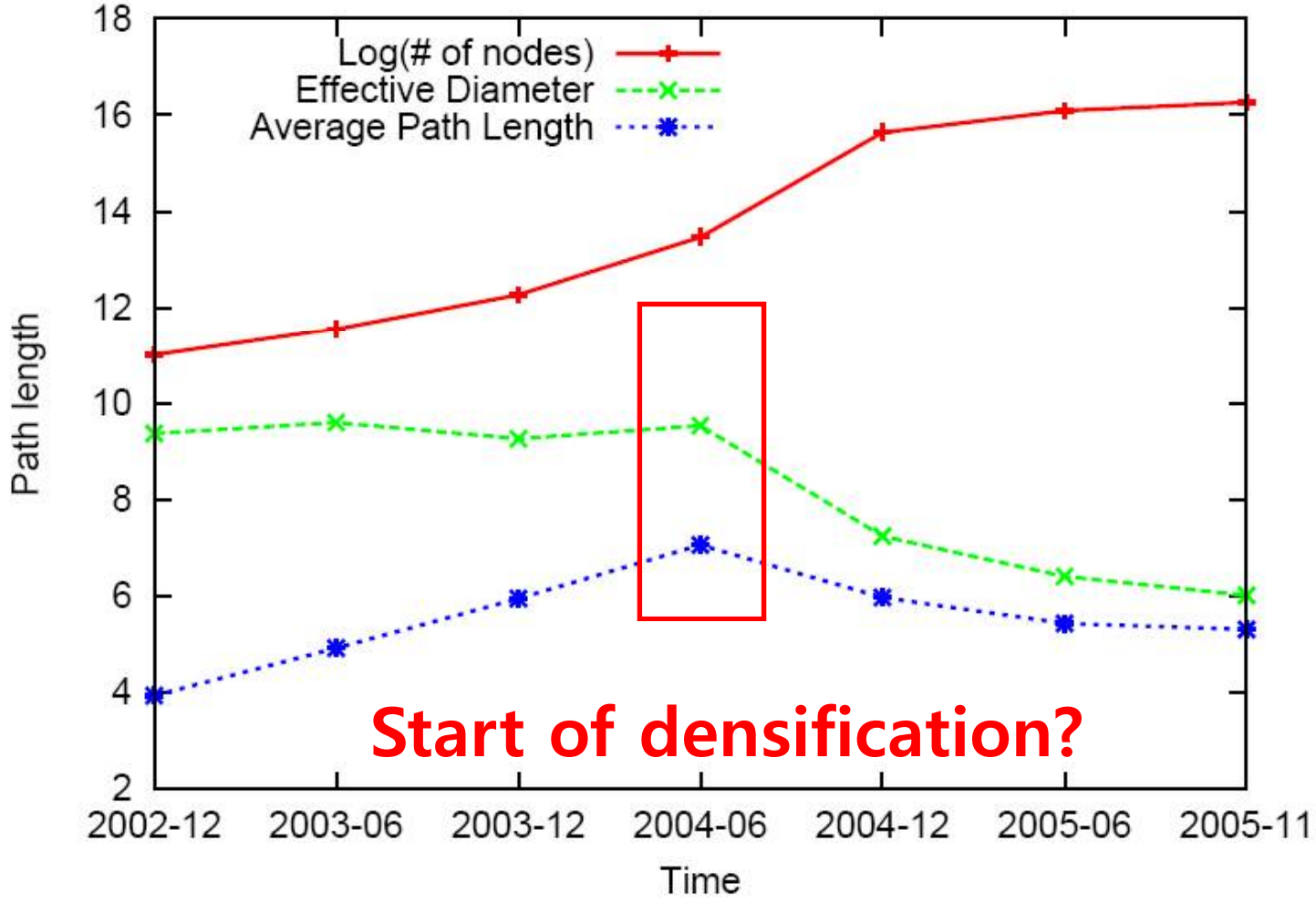


Figure 3: The effective diameter over time.

- Constant or slowly growing diameter assumption

Evolution of Path Length

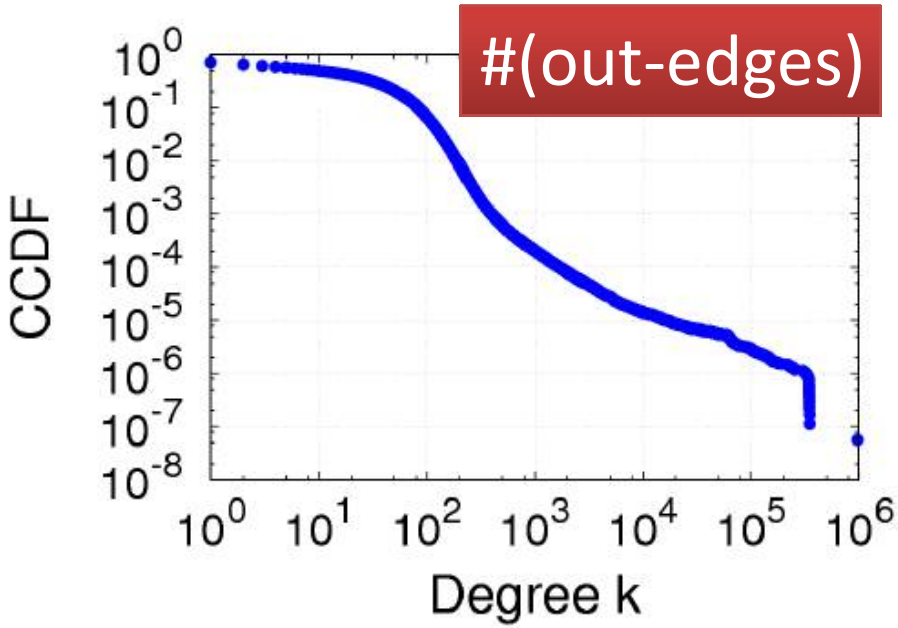


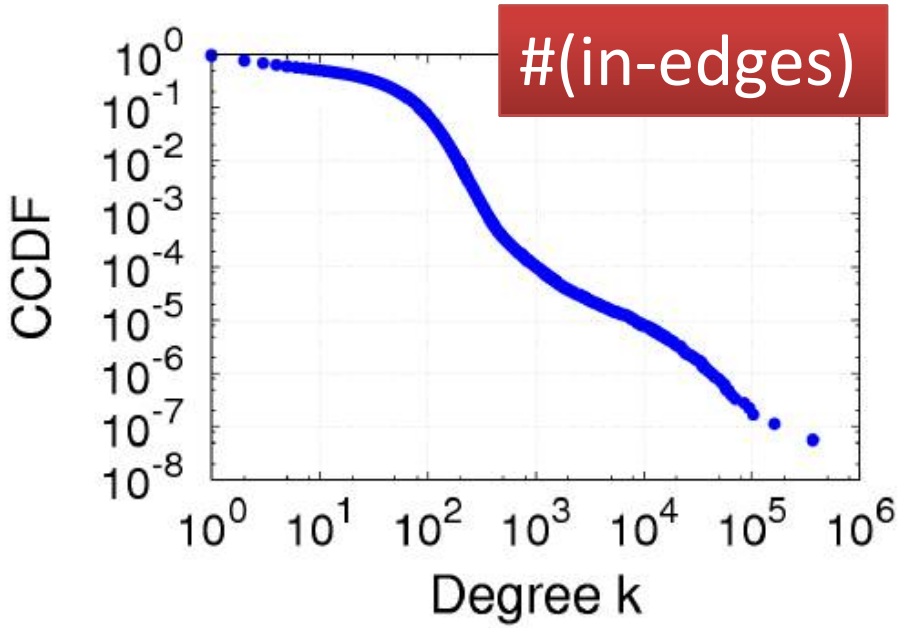
Cyworld Saturation

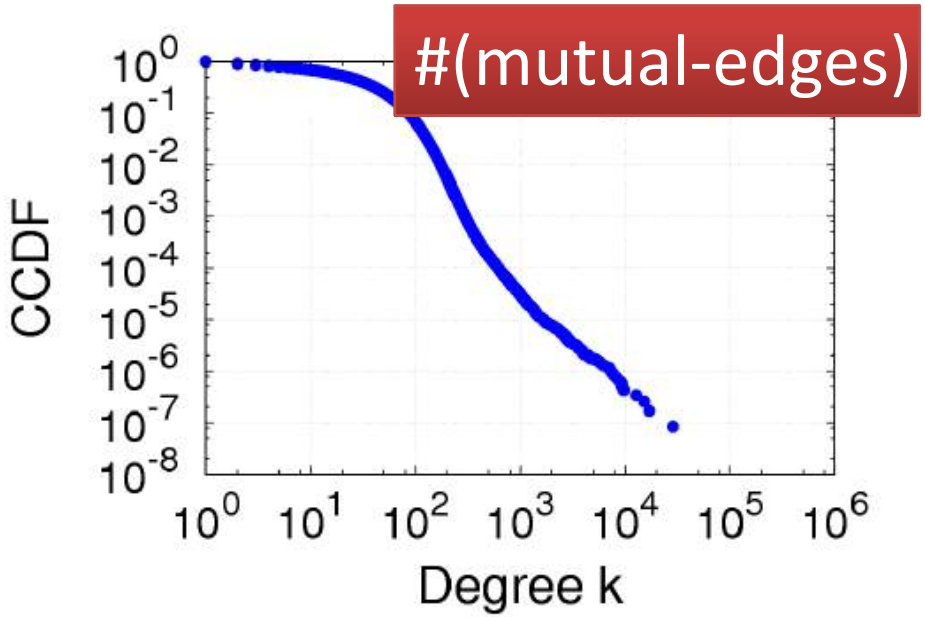
- Volume-wise
  - Exponential growth from 2003 to 2004
  - Continues to grow in 2005 and on
- Topologically
  - Network has started to “densify”
  - No network growth model fits SNS
    - Delay in ilchon building
    - Bounded population

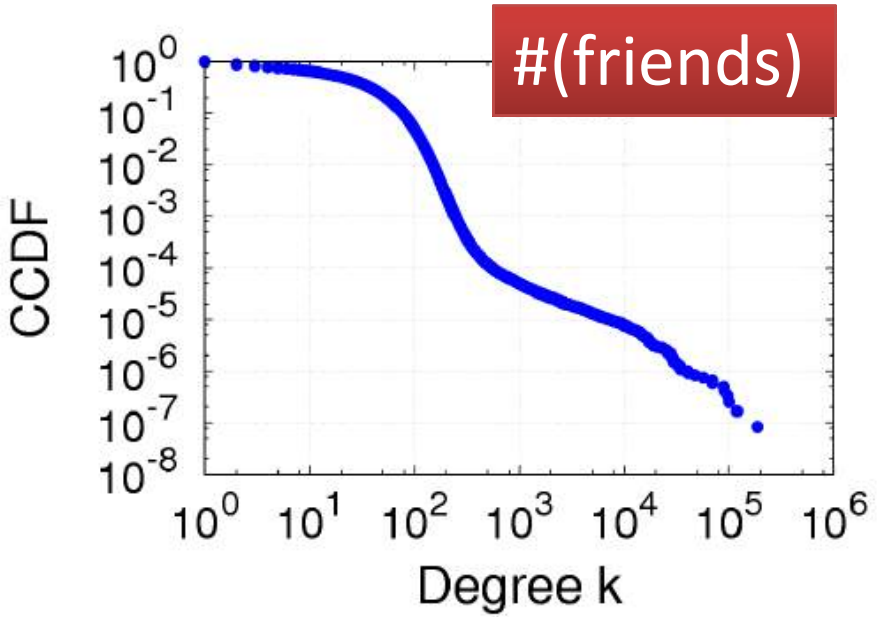
Thoughts on Evolution of SNS

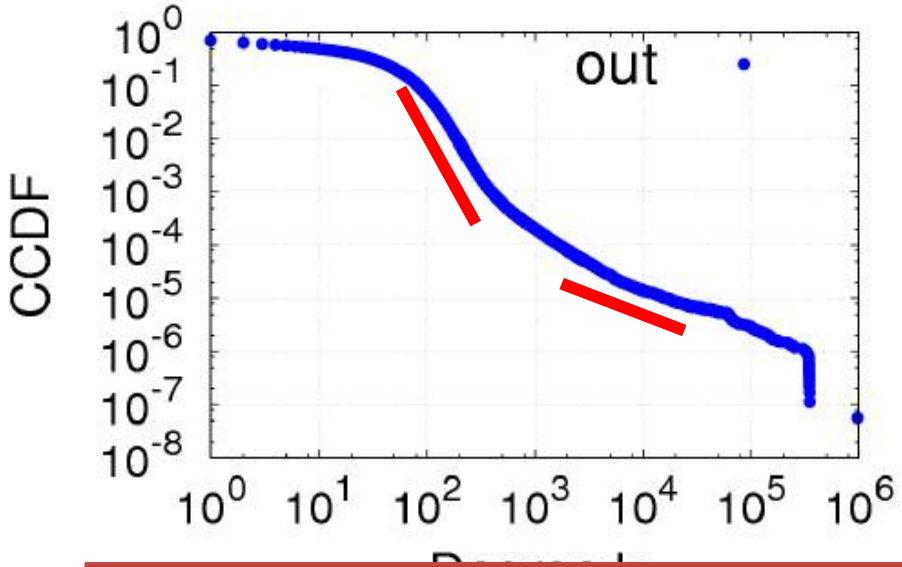
- Language barriers
- Platforms
  - Applications as “nodes” or “entity of peer pressure”

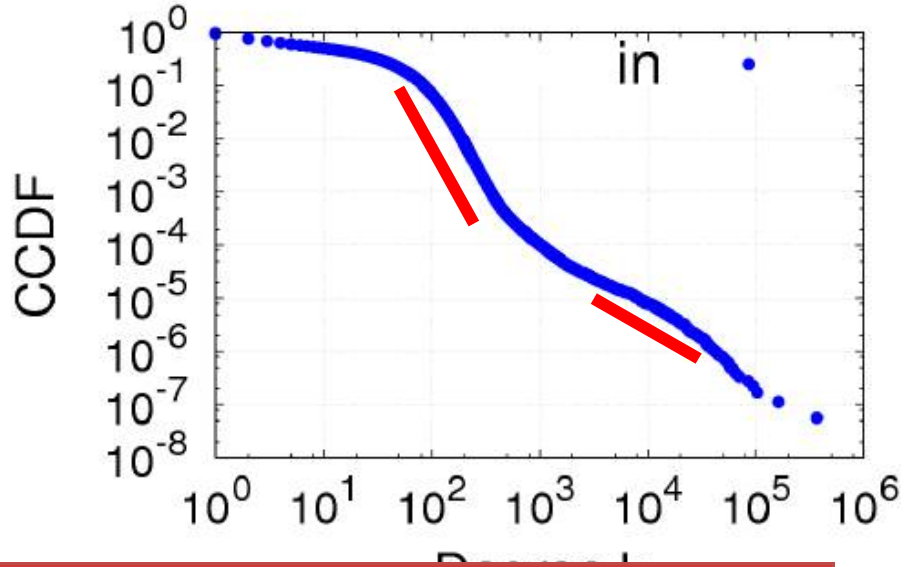

#(out-edges)
CCDF
Degree k


#(in-edges)
CCDF
Degree k

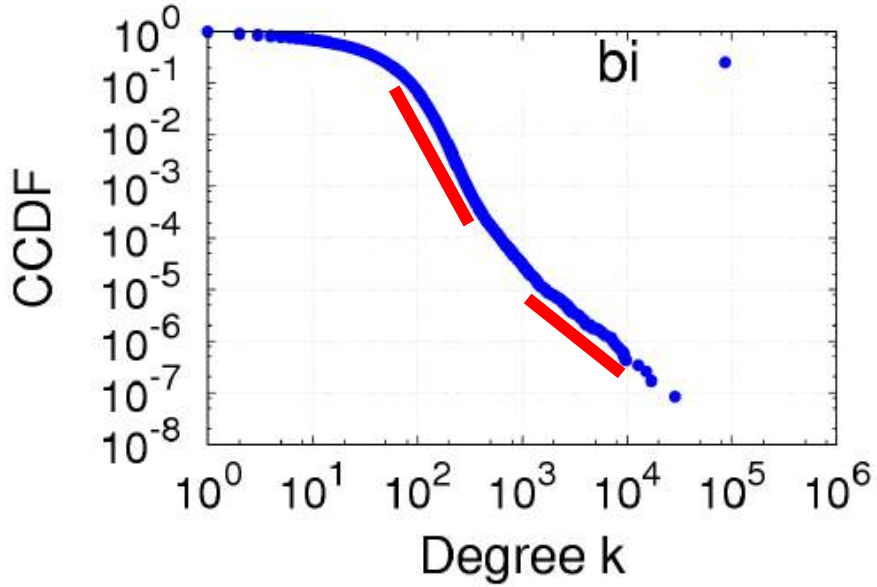

#(mutual-edges)
CCDF
Degree k

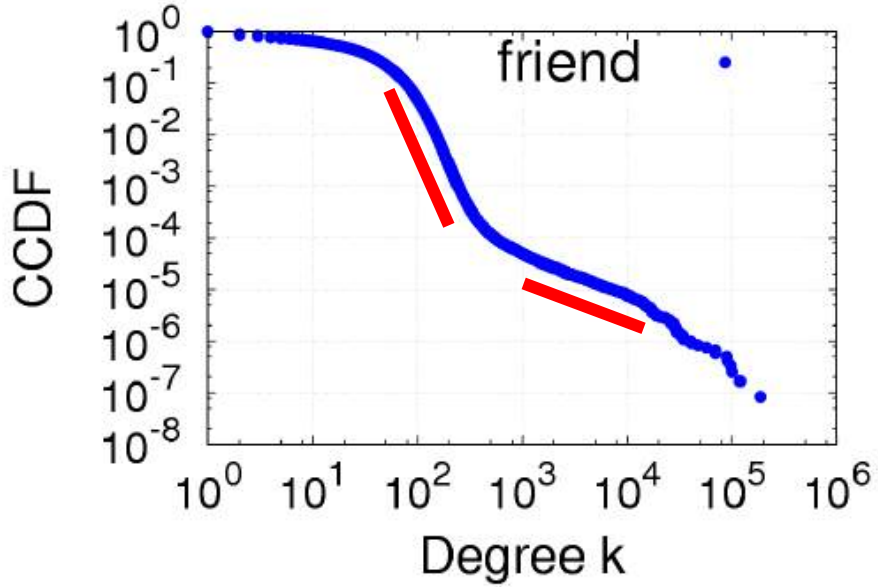

#(friends)
CCDF
Degree k

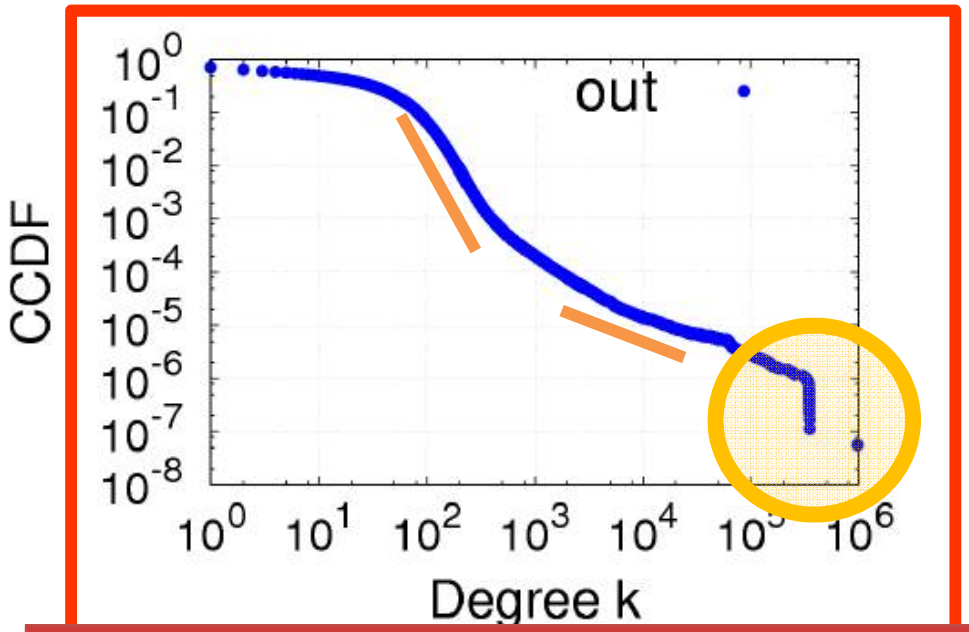


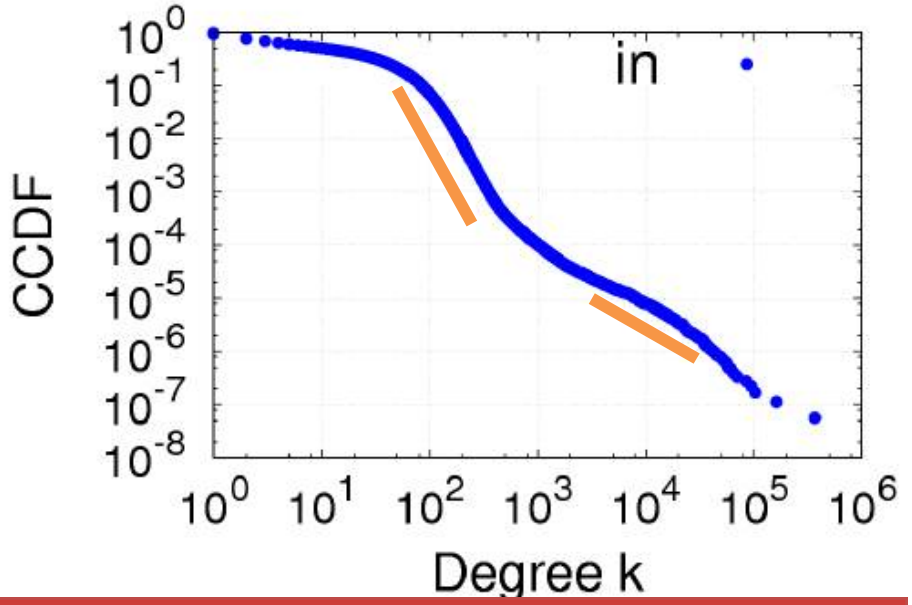


Multi-scaling behavior means heterogeneous relations

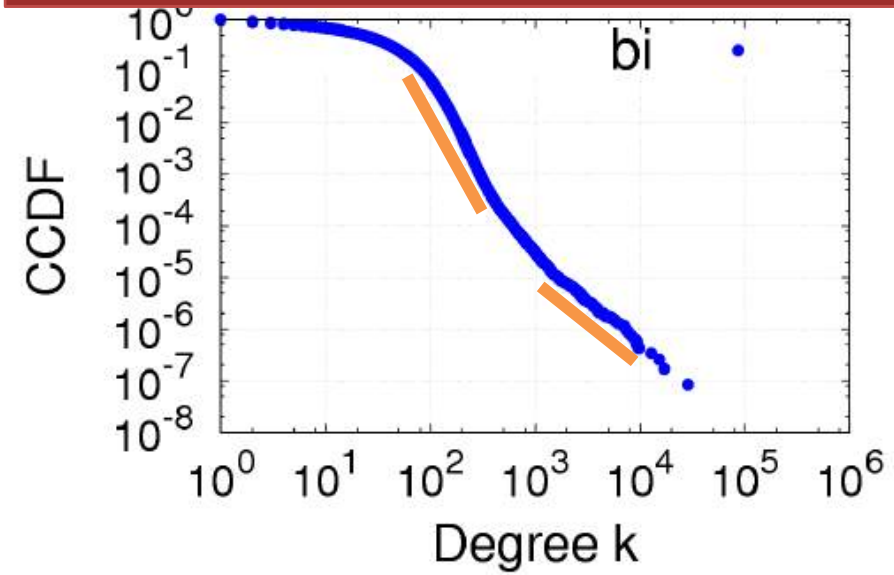


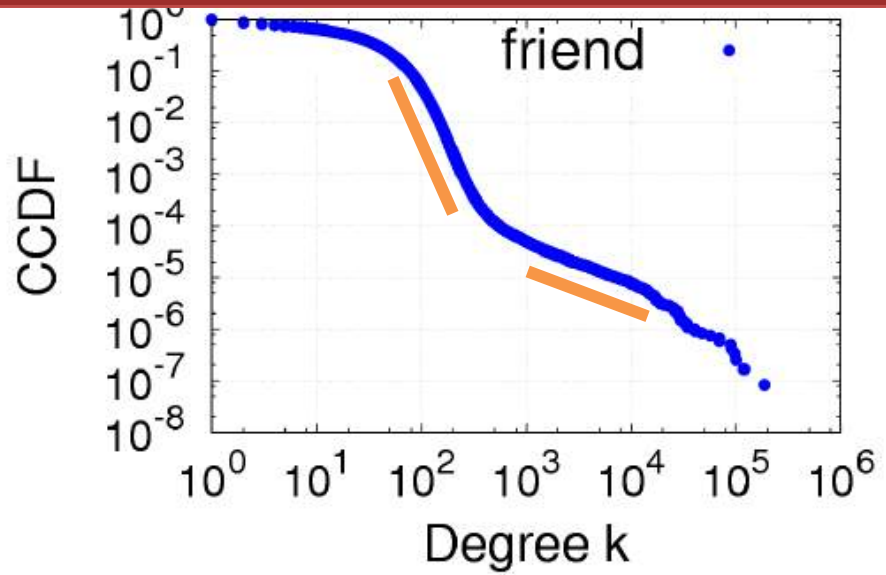


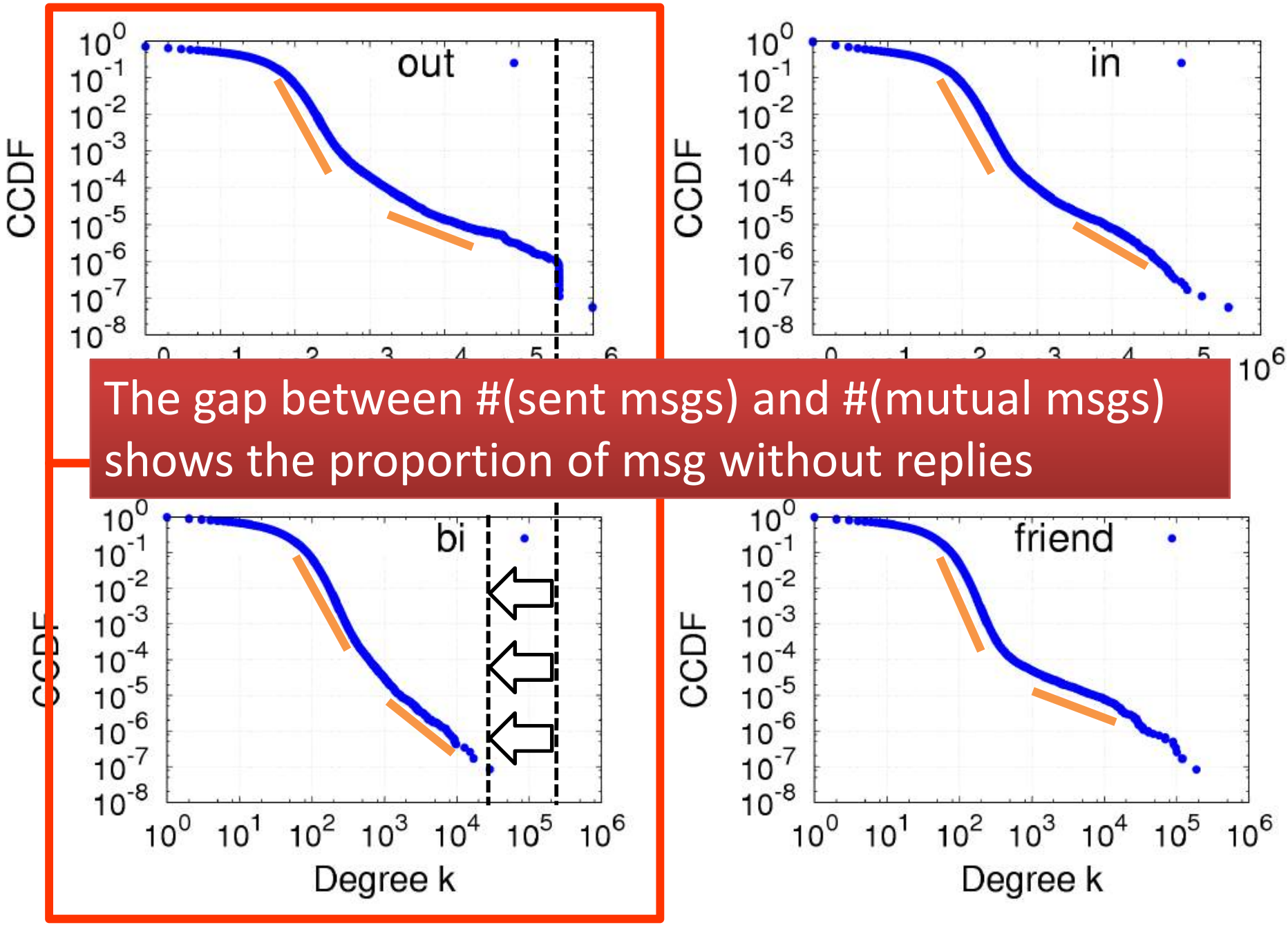




Rapid drop represents the limitation of writing capability






out
in
bi
friend
CCDF
Degree k
The gap between #(sent msgs) and #(mutual msgs) shows the proportion of msg without replies

In-depth look at Cyworld user activity

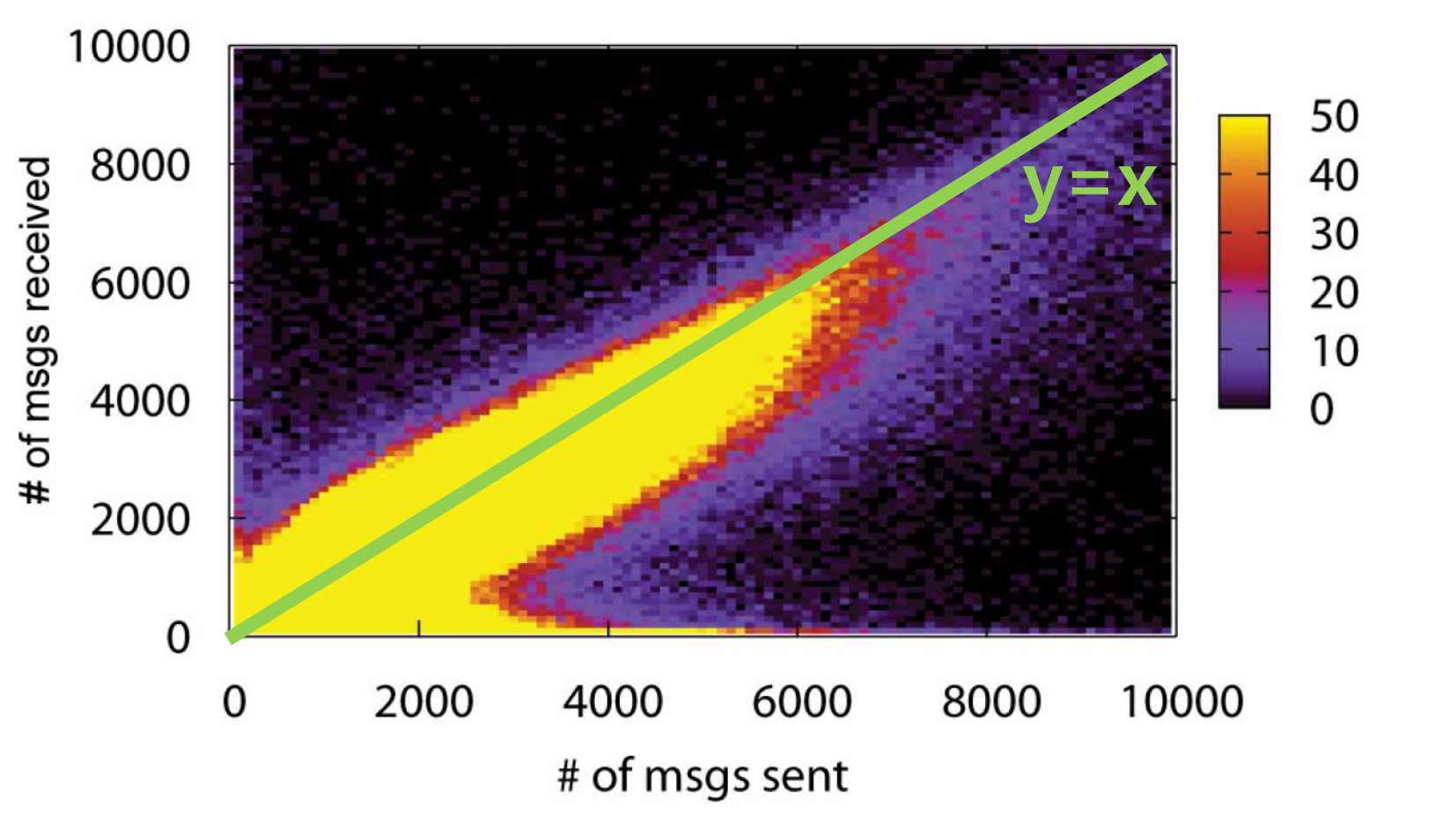
Microscopic Interaction Pattern

- Reciprocity
- Disparity
- Network motif
- Other interesting observations

Reciprocity

- Quantitative measure of reciprocal interaction
- #(sent msgs) vs. #(received msgs)

Reciprocity in user activities



Disparity

- Do users interact evenly with all friends?

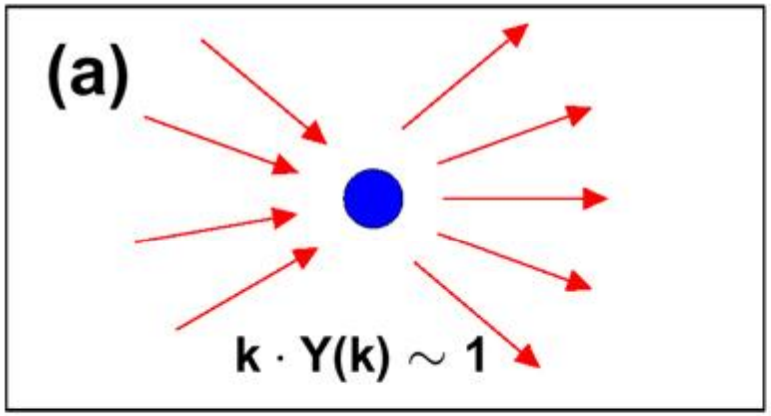
For node *i*,

$$Y(k,i) = \sum_{j=1}^{k} \left\{ \frac{w_{ij}}{\sum_{l=1}^{k_{in}} w_{li}} \right\}^2$$

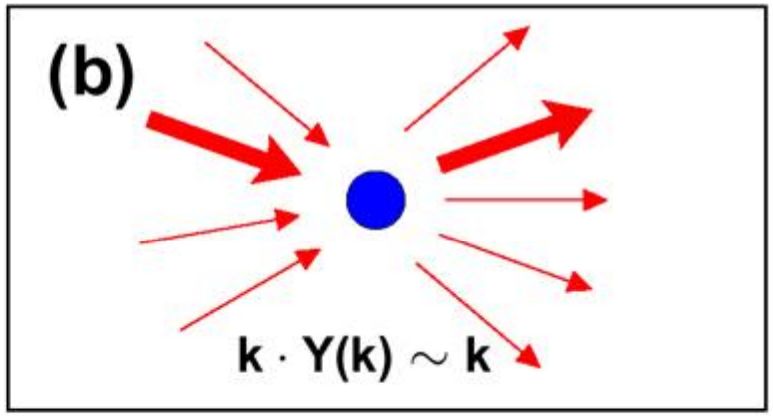
Y(k) is average over all nodes of degree *k*

Journal of Physics A: Mathematical and General, 20:5273–5288, 1987.

Interpretation of Y(k)

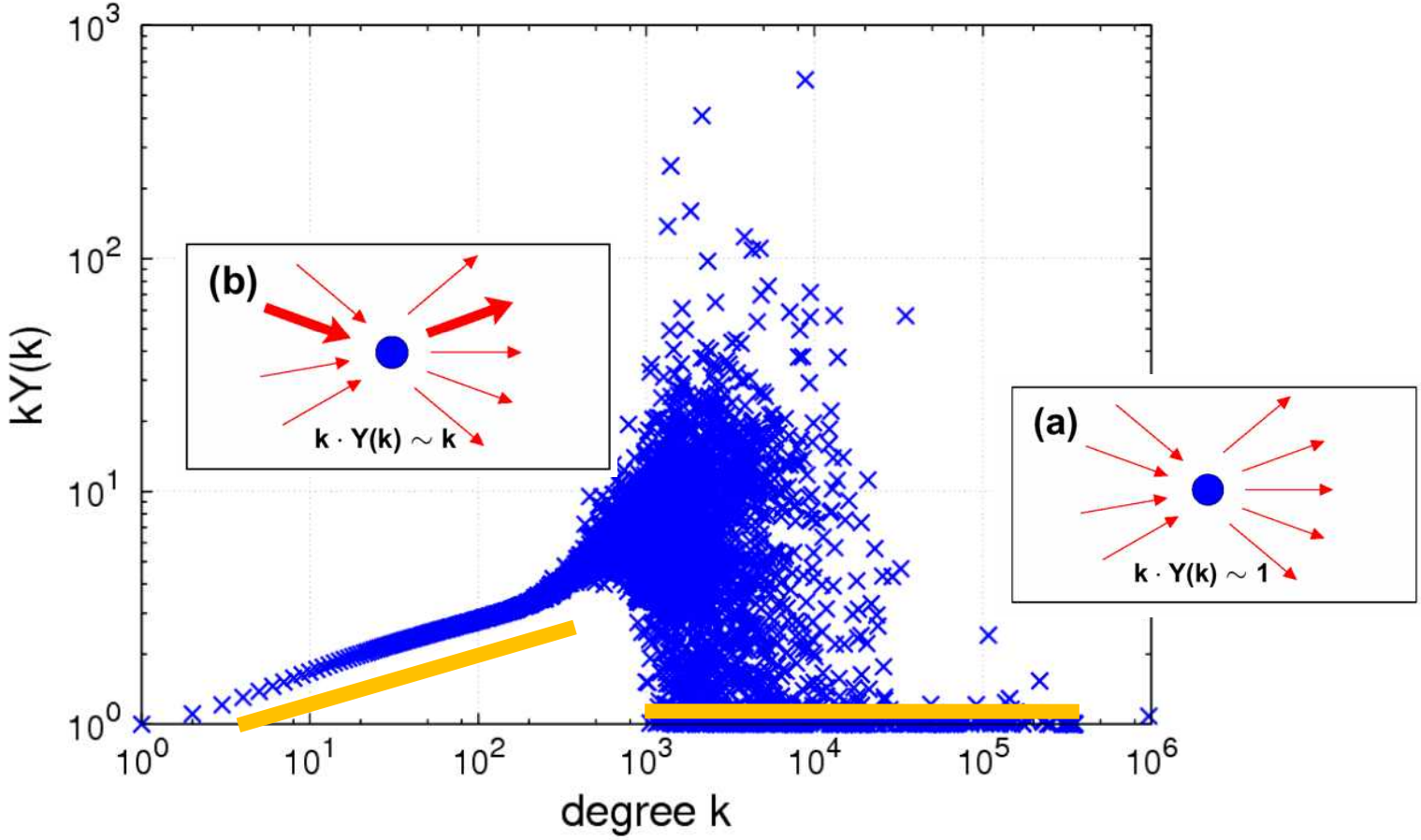


Communicate evenly



Have dominant partner

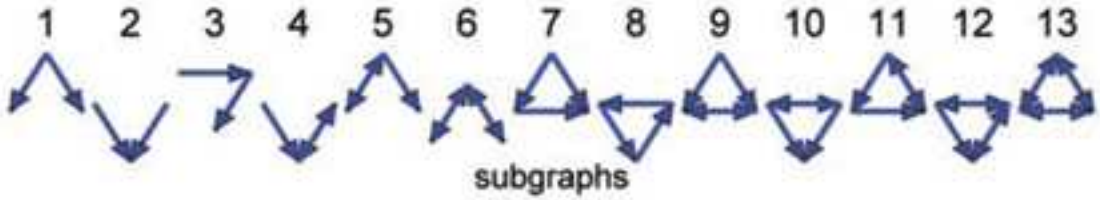
Disparity in user activities



Network Motifs

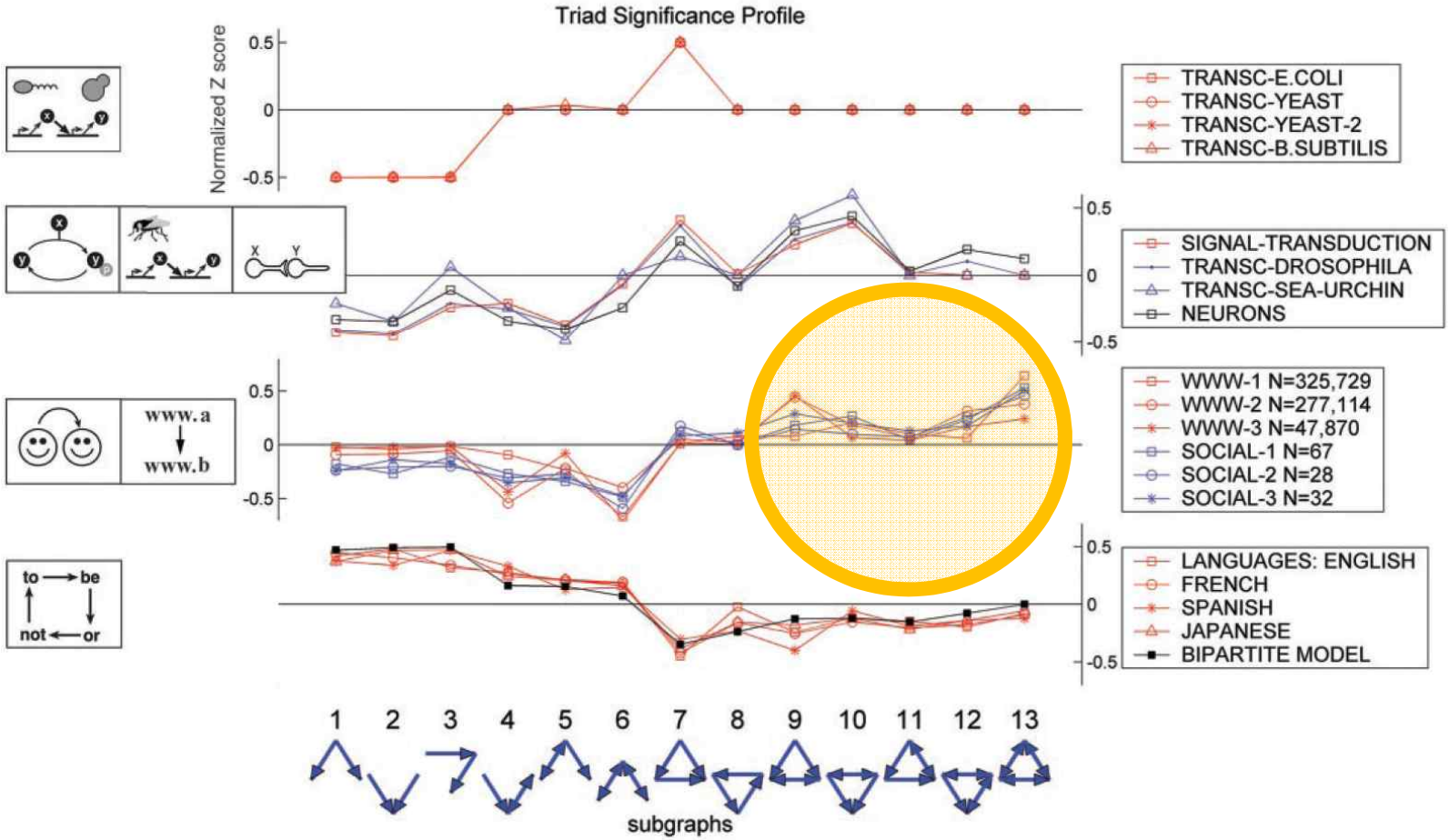
Science, Vol. 298, 824-827

- All possible interaction patterns with 3 users



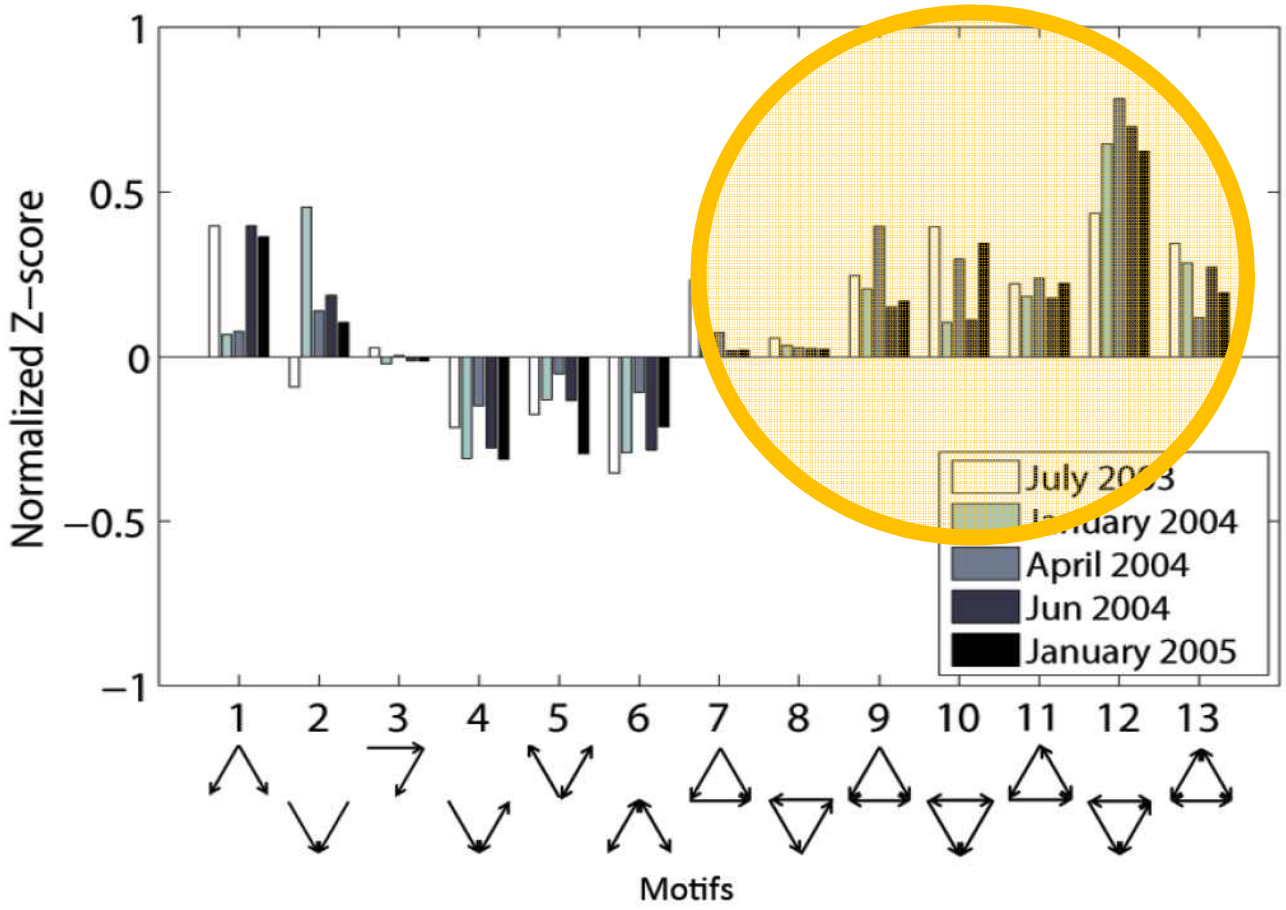
- Proportions of each pattern (motif) determine the characteristic of the entire network

Motif analysis in complex networks

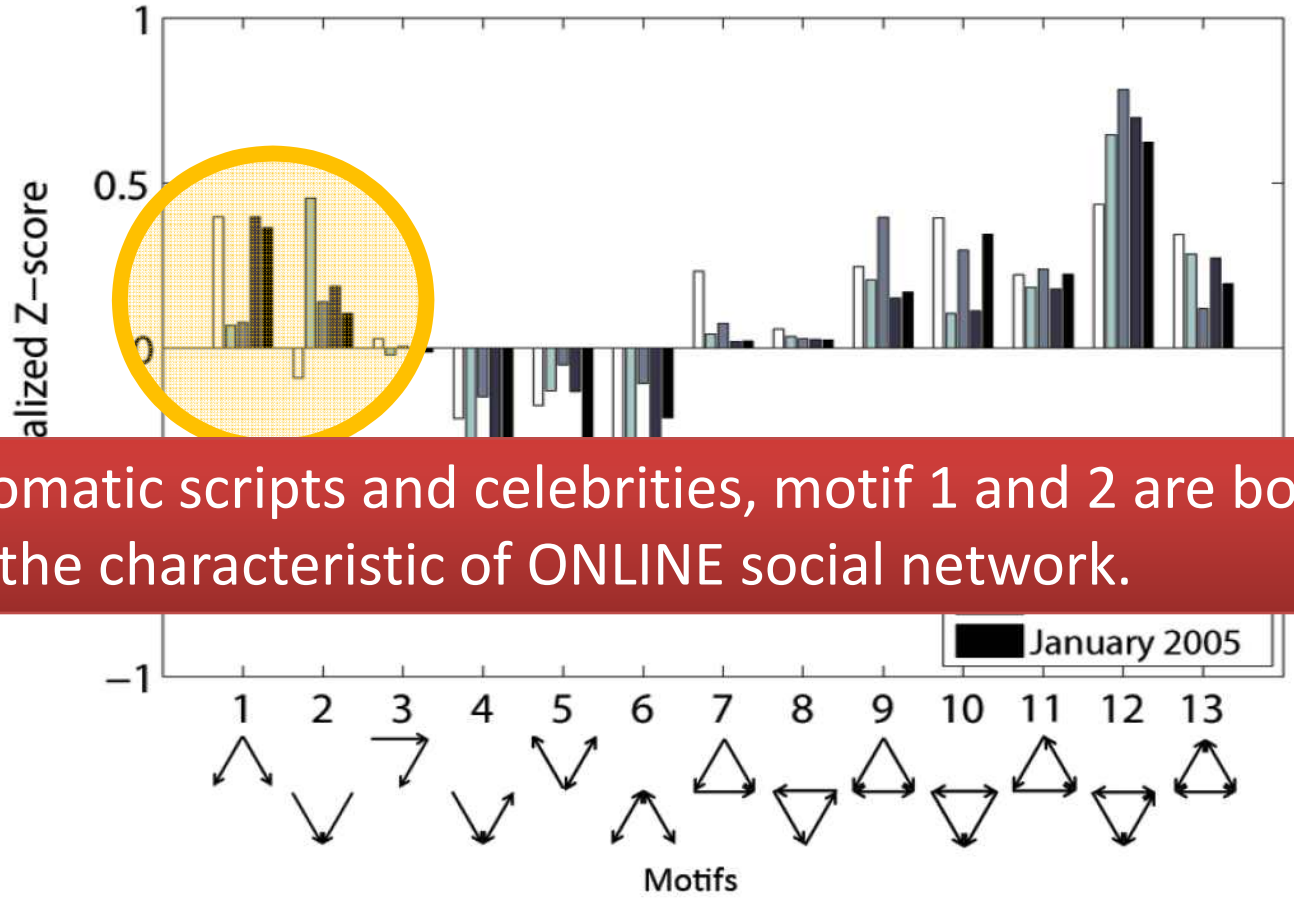


Science, Vol. 303, no. 5663, pp 1538-1542, 2004

Network motifs in user activities

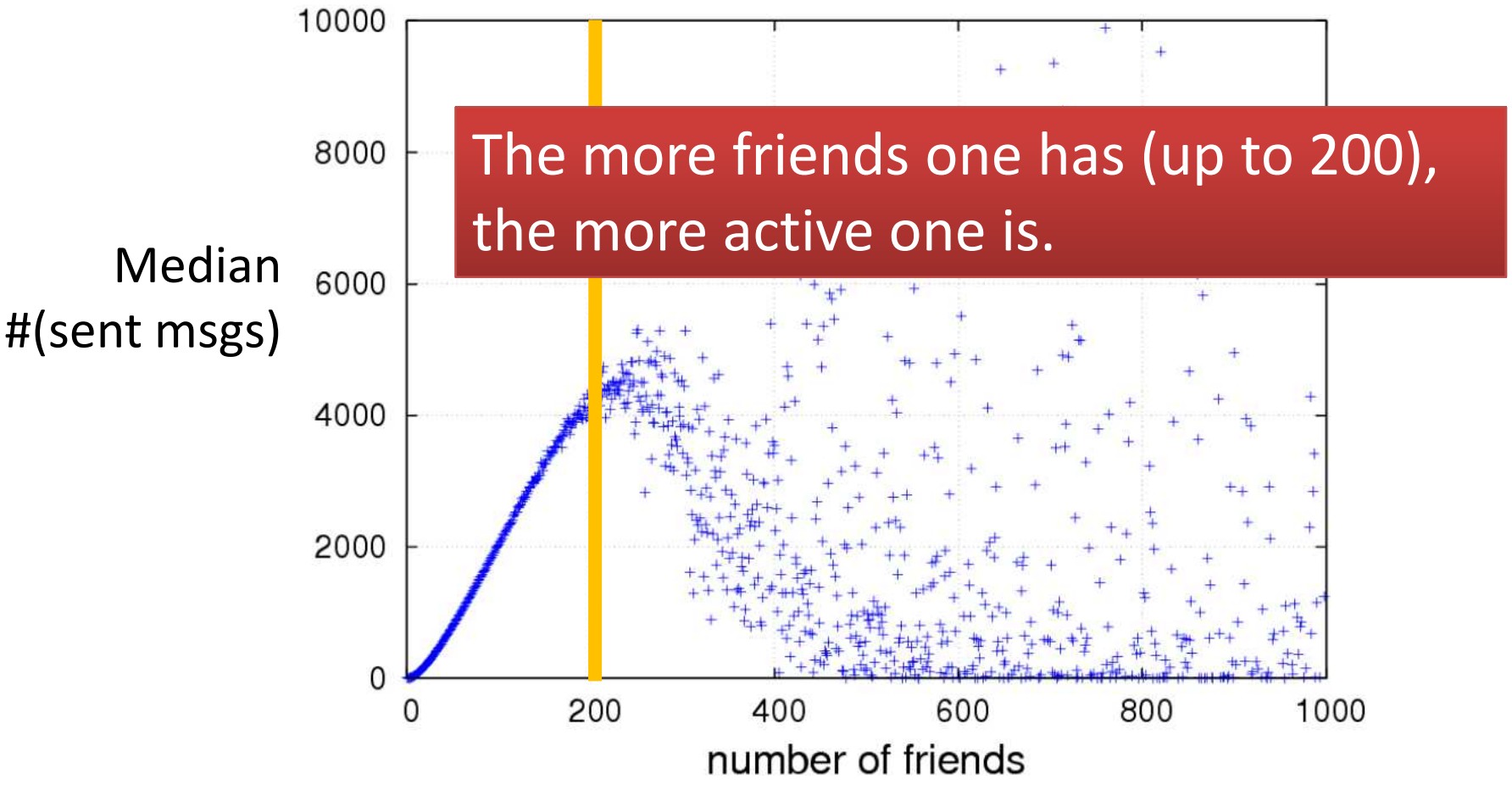


Network motifs in user activities

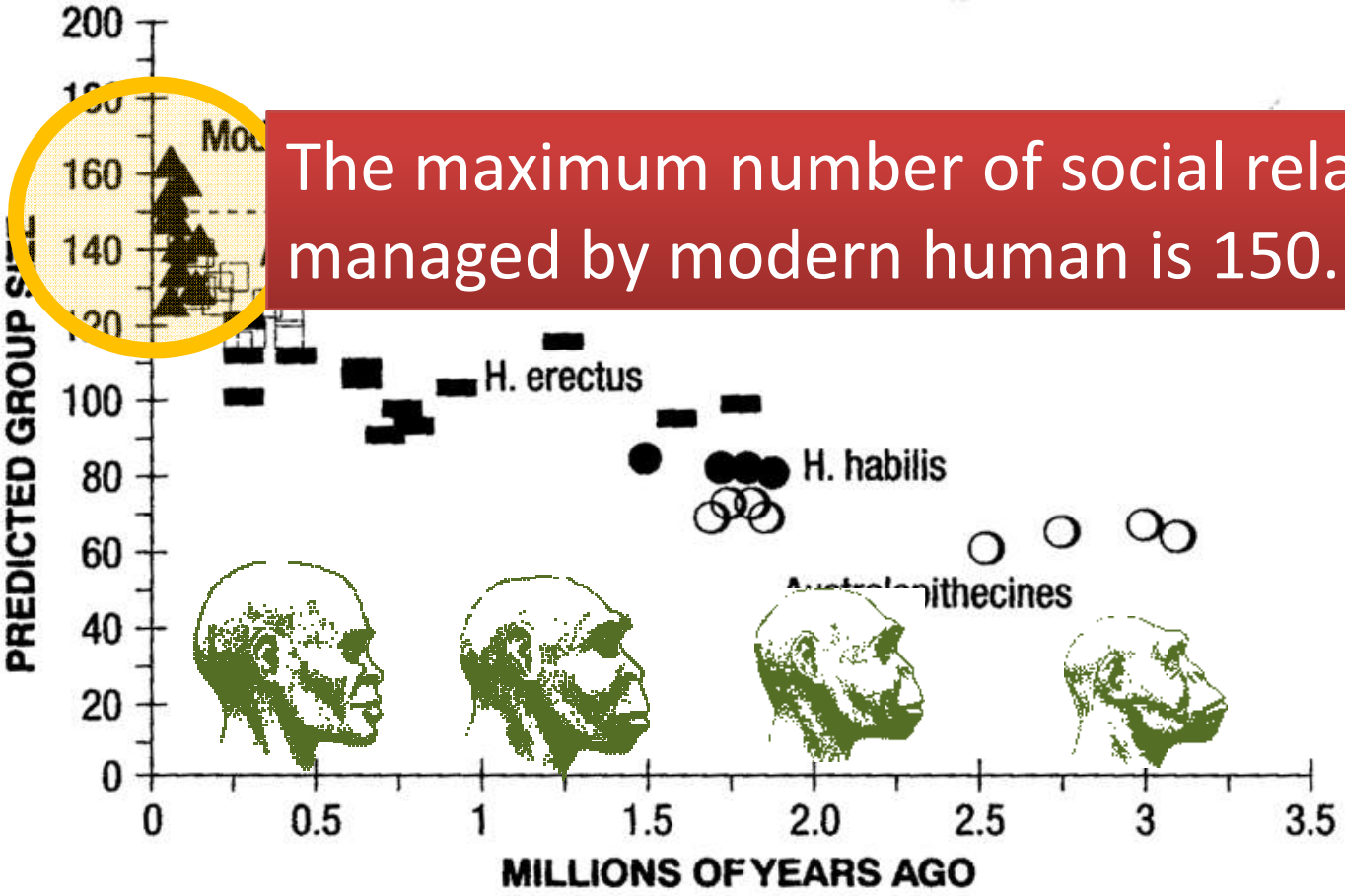


By automatic scripts and celebrities, motif 1 and 2 are boosted.
This is the characteristic of ONLINE social network.

#(friends) stimulate interaction?



Dunbar's number



The maximum number of social relations managed by modern human is 150.

Behavioral and brain scineces, 16(4):681–735, 1993

200 vs. 150

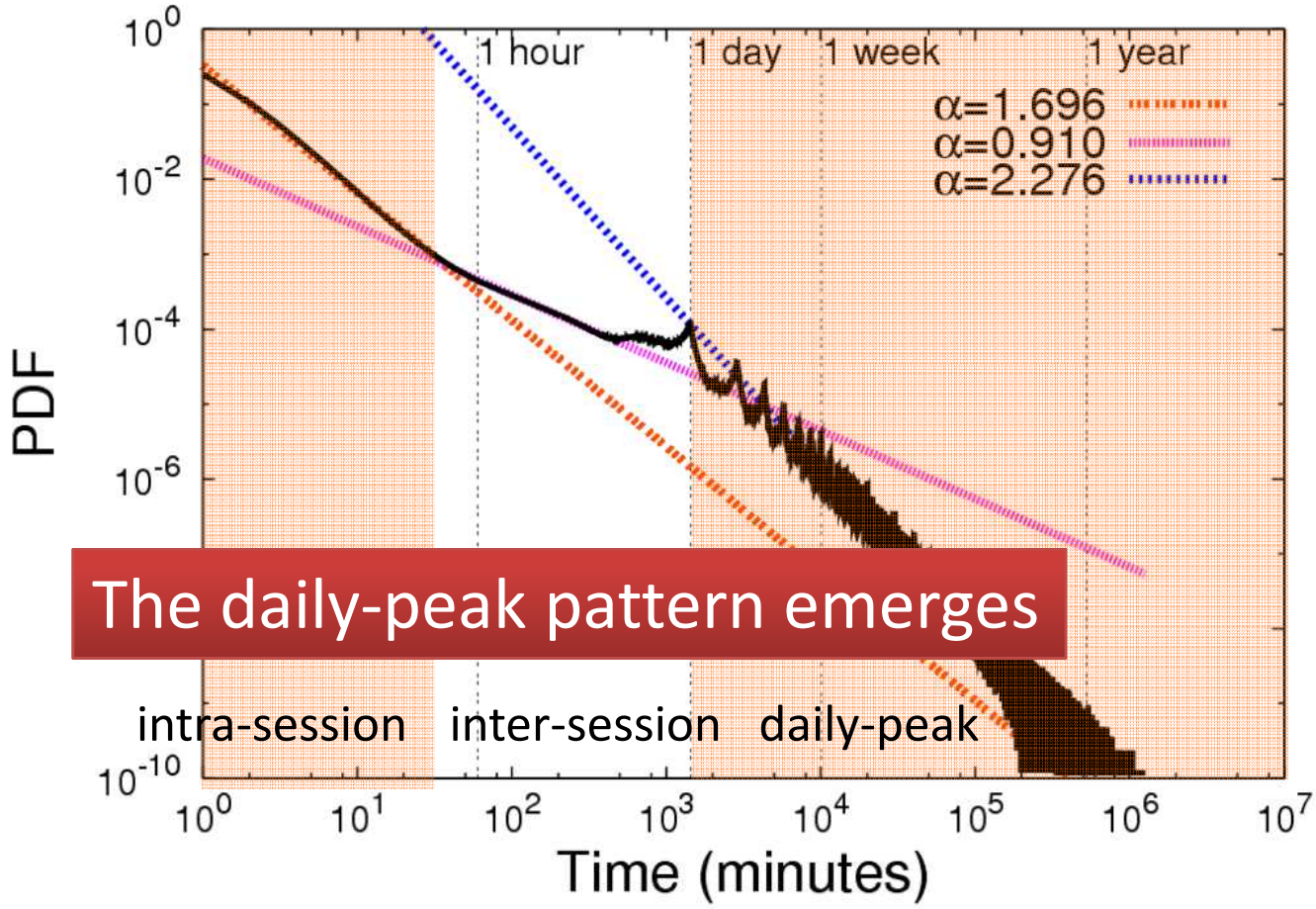
- Does human's capacity really enlarge?
  - Yes, technology help users to manage relations
  - No, it is the only inflation #(friends)

Time interval between msgs

- Is there a particular pattern in writing a msg?

- Bursts in human dynamics
  - e-mail
  - MSN messenger

Time interval between msgs



The daily-peak pattern emerges

Nature, 435:207–211, 2005
Proceedings of WWW2008, 2008

Part III:
Consistent Community Identification

Community identification

- Our goal: mine "implicit" groups
- Metric of interest = modularity

$$Q = \sum_i (e_{ii} - a_i^2)$$

$\mathbf{e_{ii}}$ is # links within comm/all links

$\mathbf{a_{ii}}$ is # link crossing the boundary/all links

Problem Statement

How Can We Identify
Consistent Communities
Independent of Input Ordering?

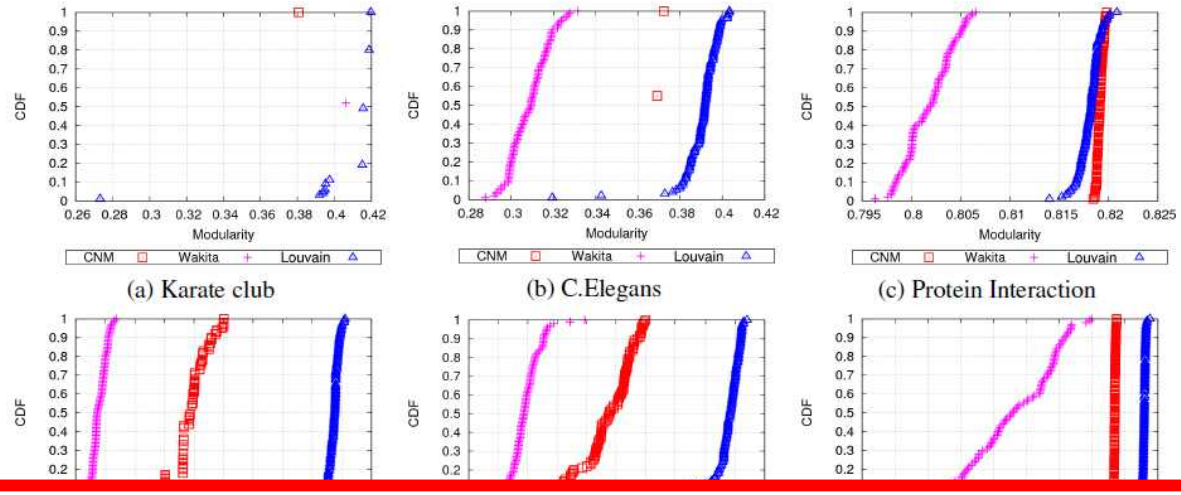
Three Algorithms

- CNM
  - Start with a node per community
  - Merge two communities that maximize ΔQ
  - One big group by another
- Wakita
  - Balancing act in community merging
- Louvain
  - Aggressive merging at each ΔQ

8 Data Sets

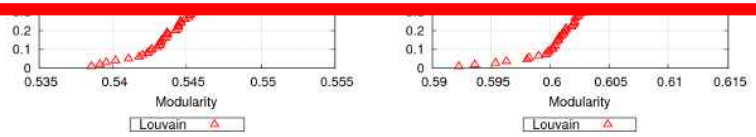
- Karate club
- C.Elegans
- Protein
- BBS
- AS graph
- WWW
- Wikipedia
- CyWorld

Modularity



(a) Karate club (b) C.Elegans (c) Protein Interaction

Modularity values all different in every run



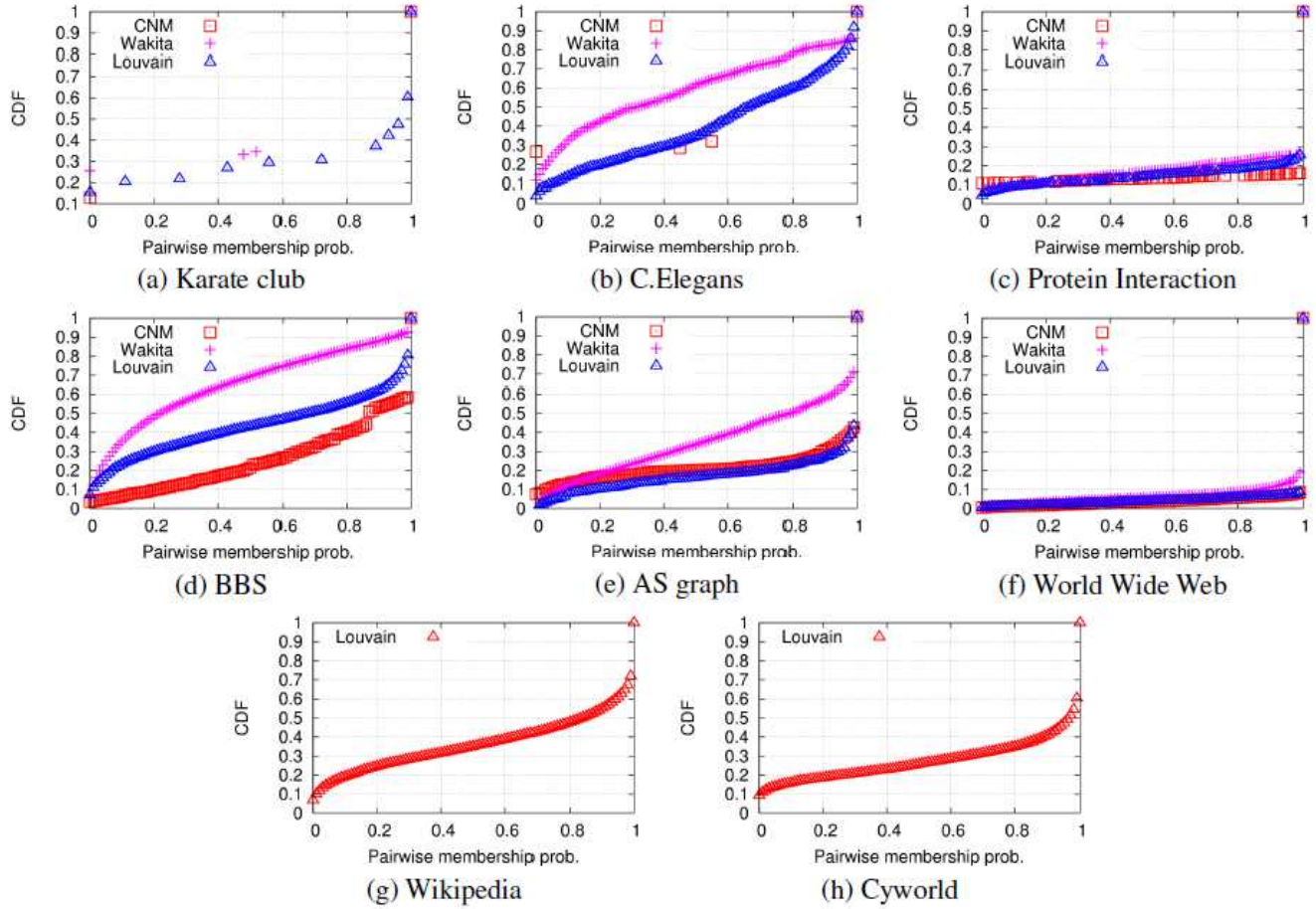
(g) Wikipedia (h) Cyworld

Metrics of “consistency”

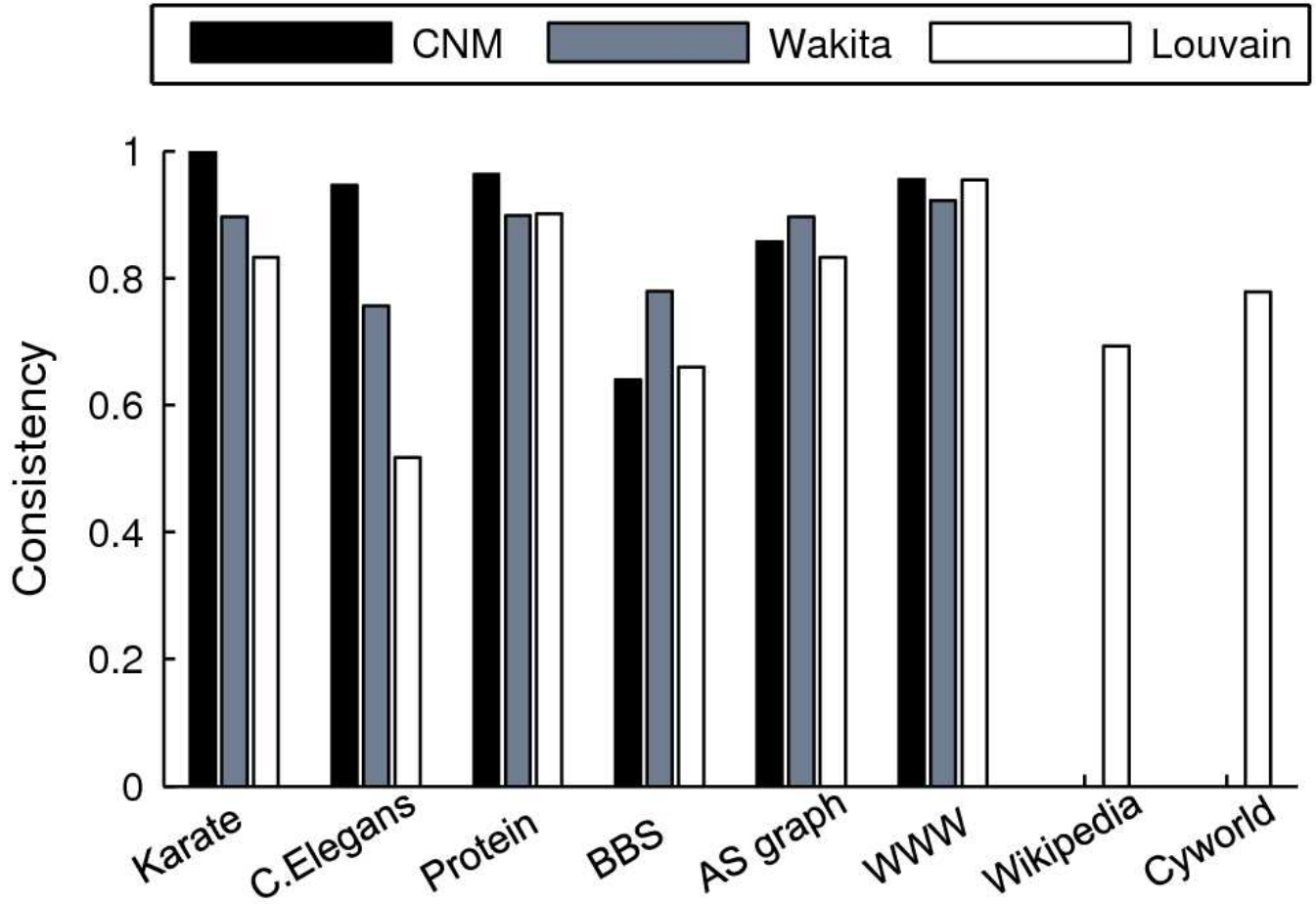
- Pairwise membership probability

- Consistency

Pairwise Membership Probability



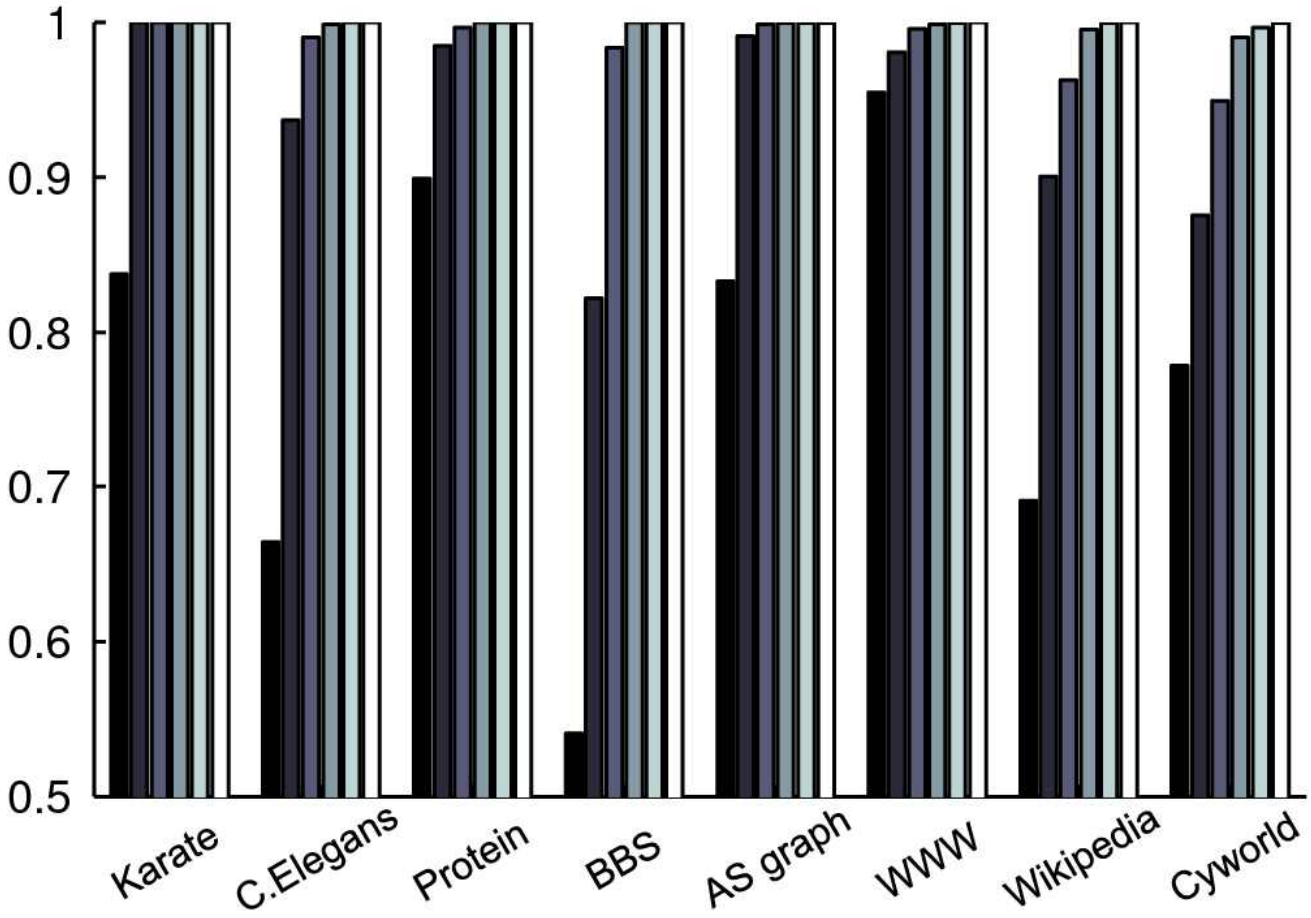
Consistency



Our algorithm

- Run any algorithm for 100 times
- Compute pairwise membership probabilities
- Use the probabilities as link weights
- Run the algorithm again for 100 times
- Repeat above until consistency reaches 1

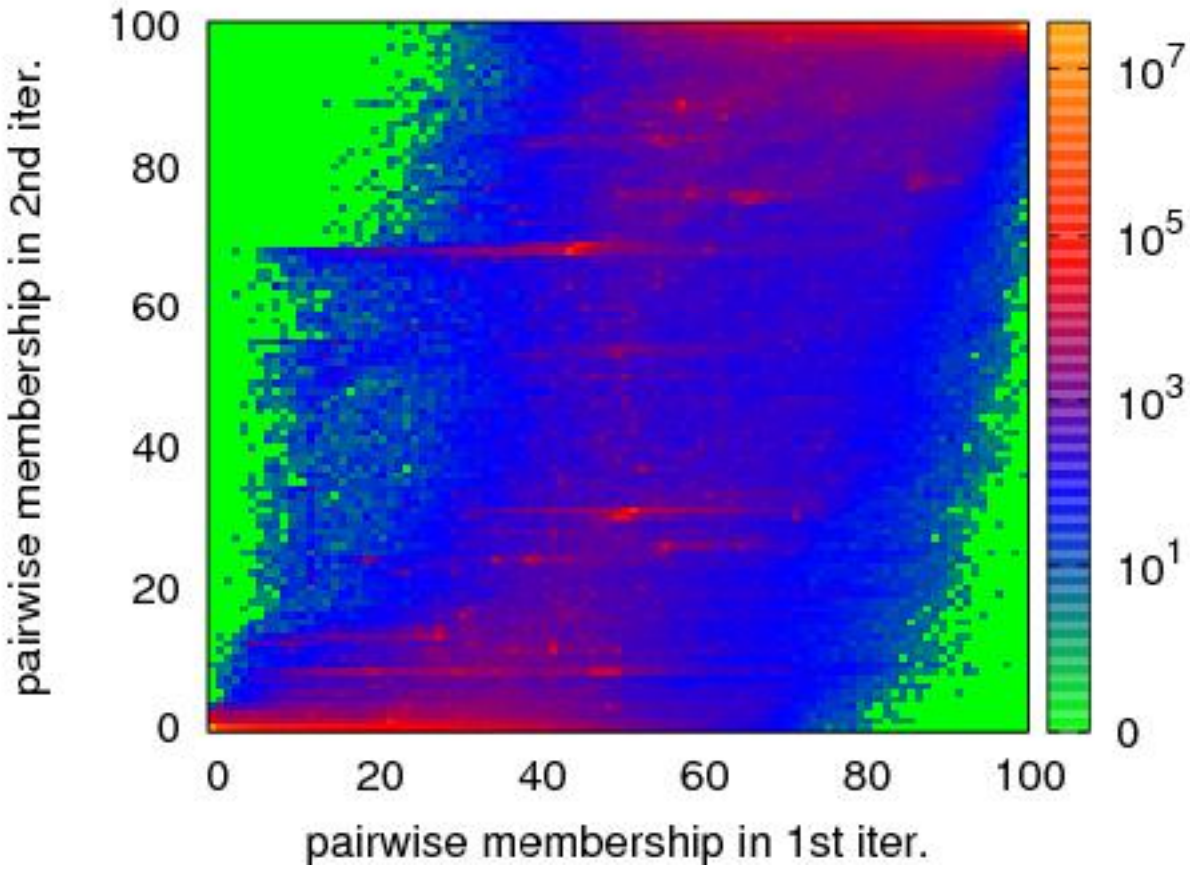
Convergence of Consistency in 5 Cycles



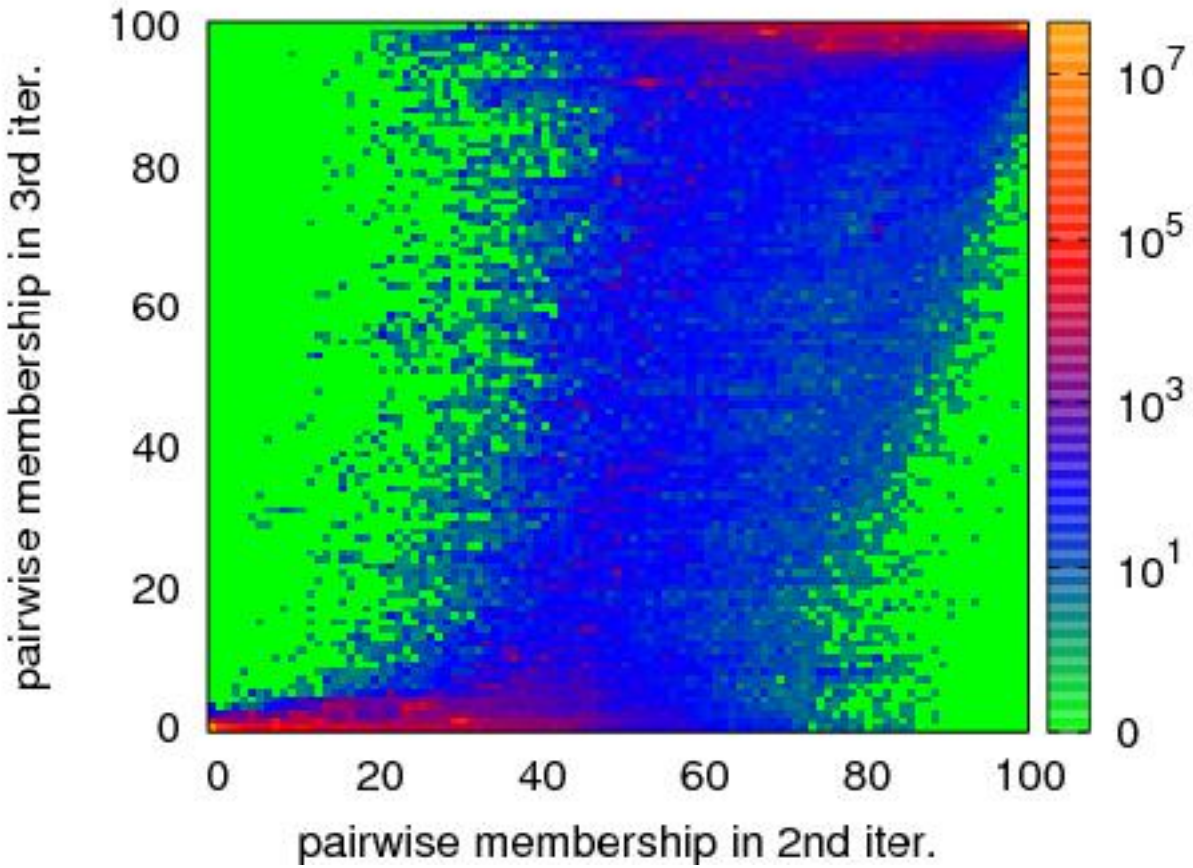
Preliminary Analysis of the AS graph

- Convergence to consistency
- Community Size
- Jaccardi Index
- A Close Look at Some Communities

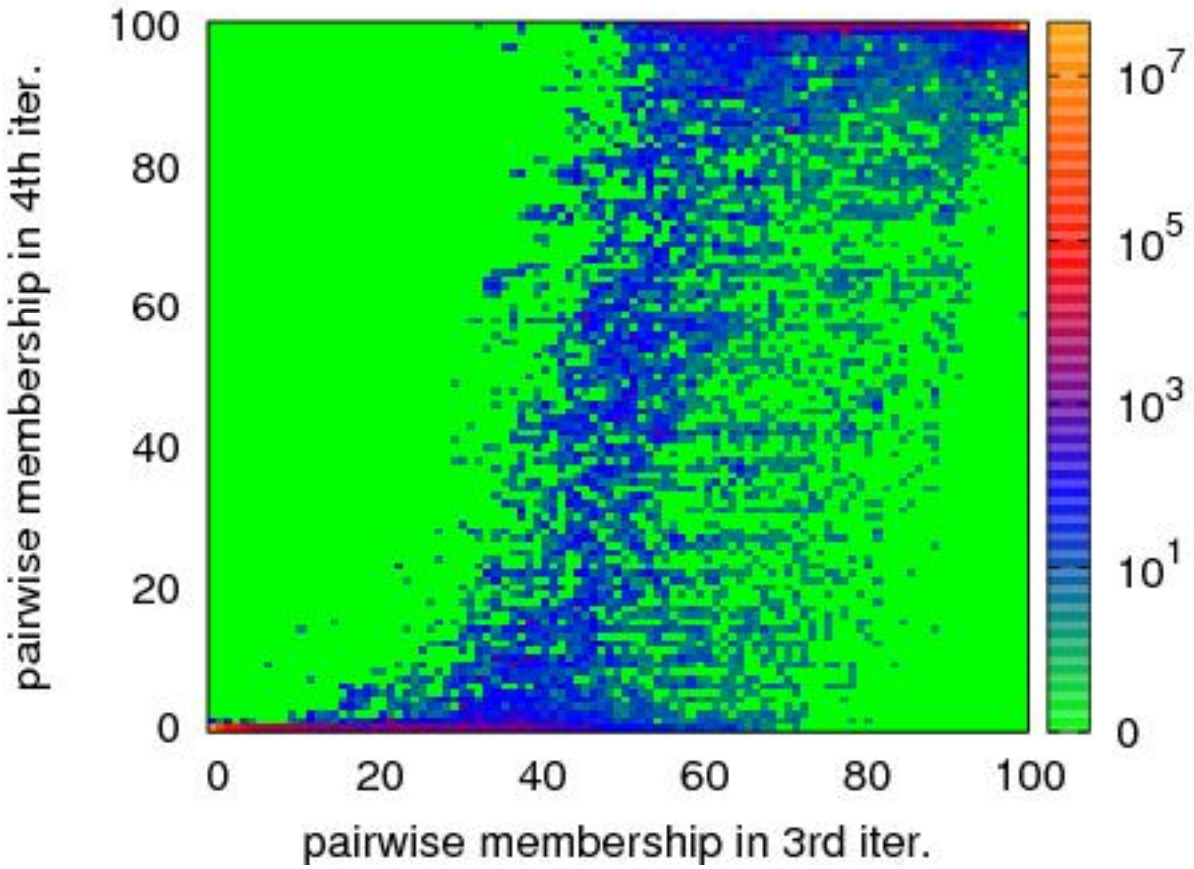
Convergence to consistency



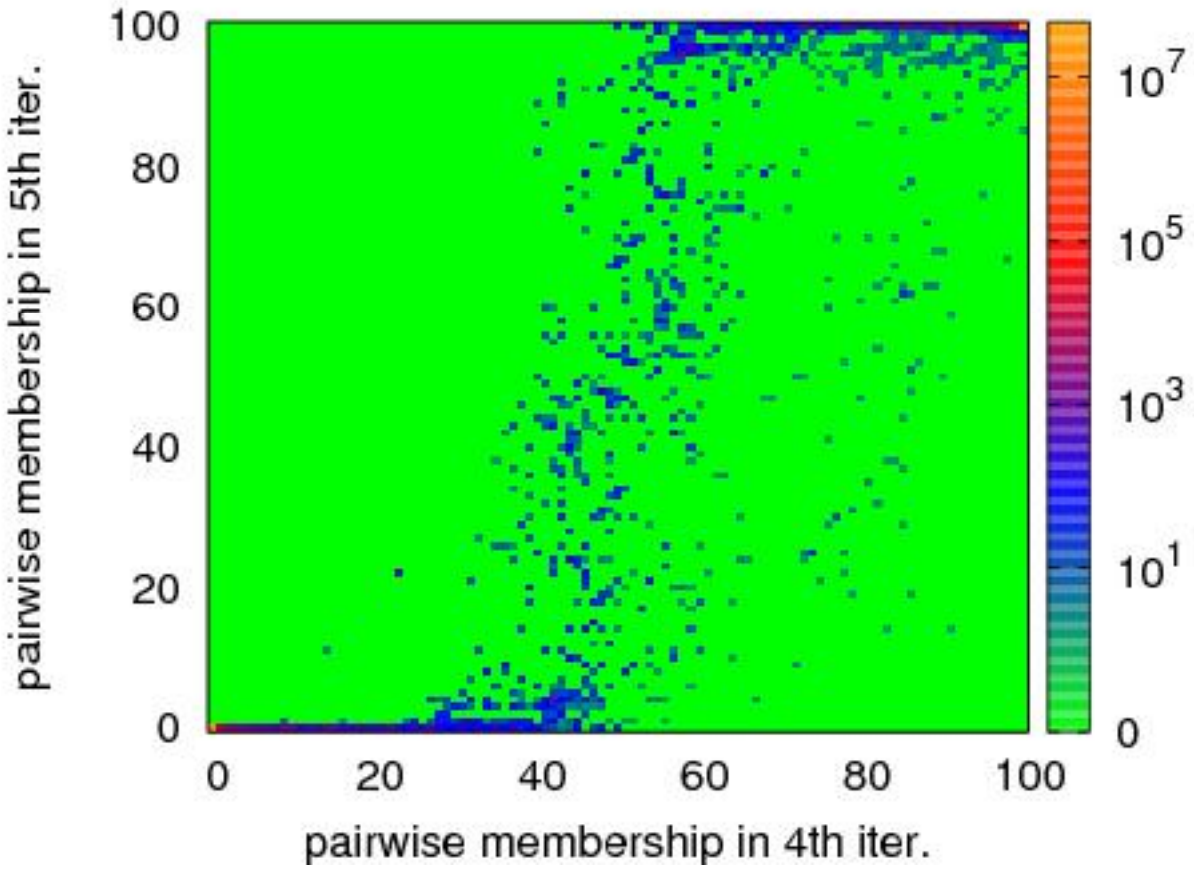
Convergence to consistency



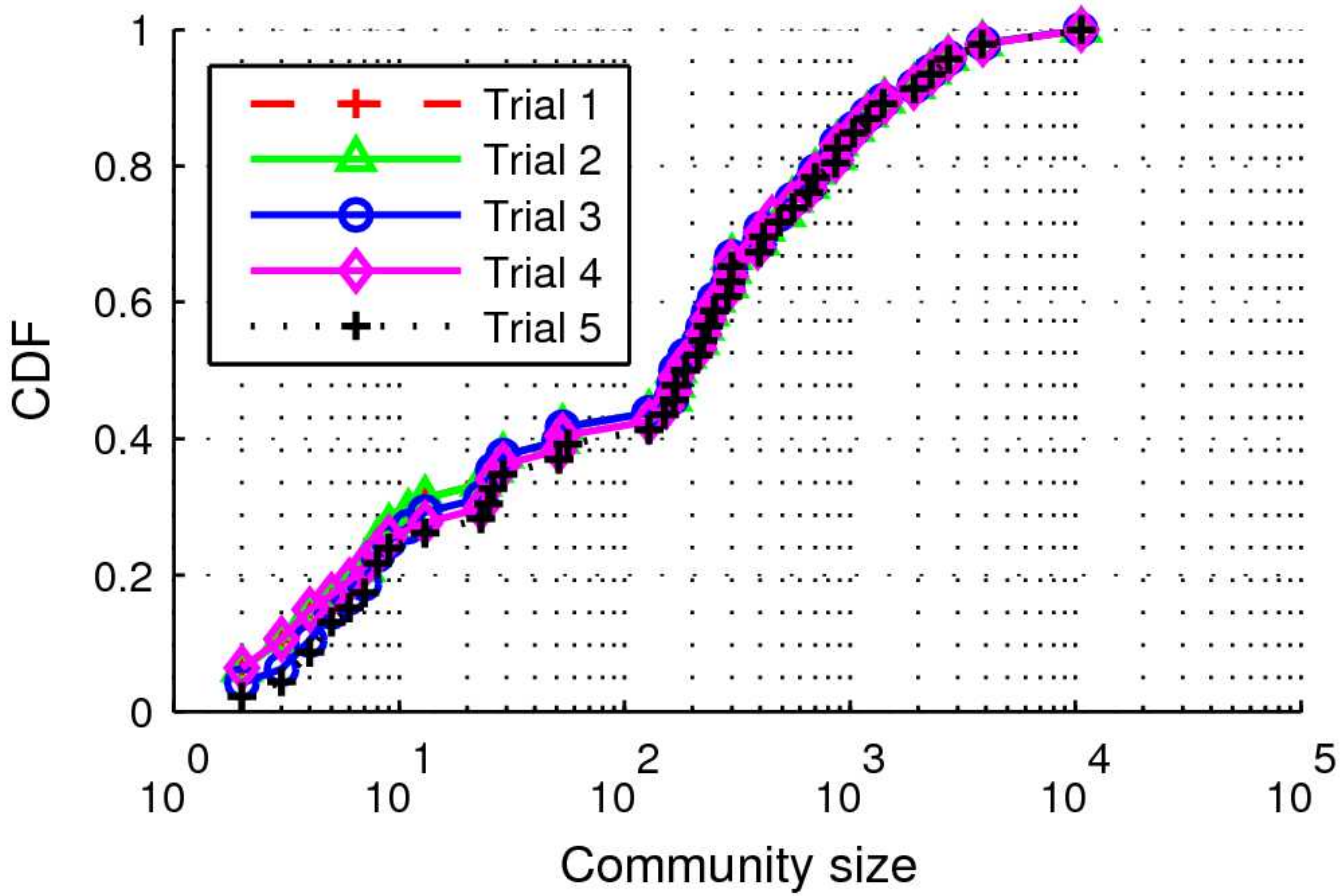
Convergence to consistency



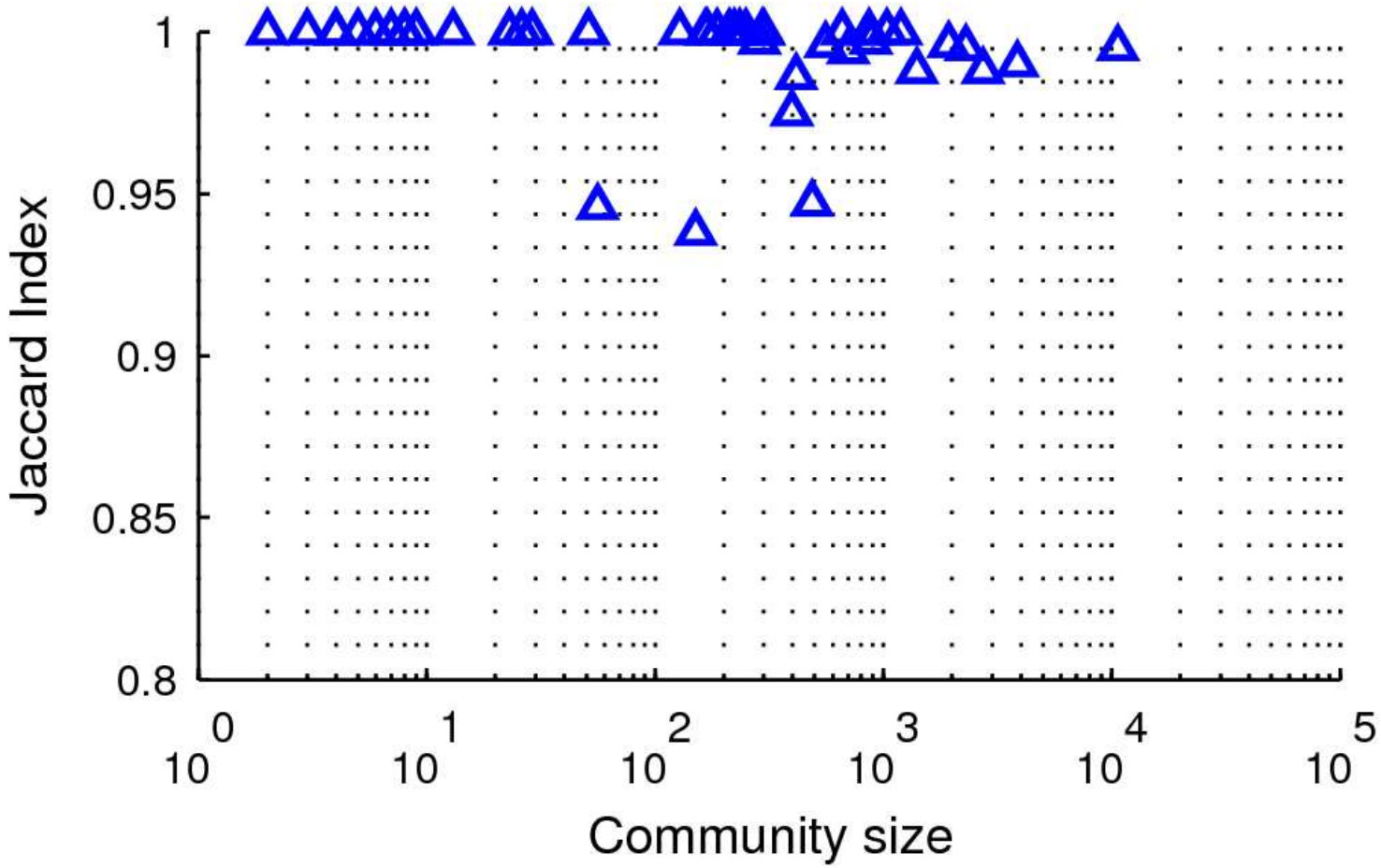
Convergence to consistency



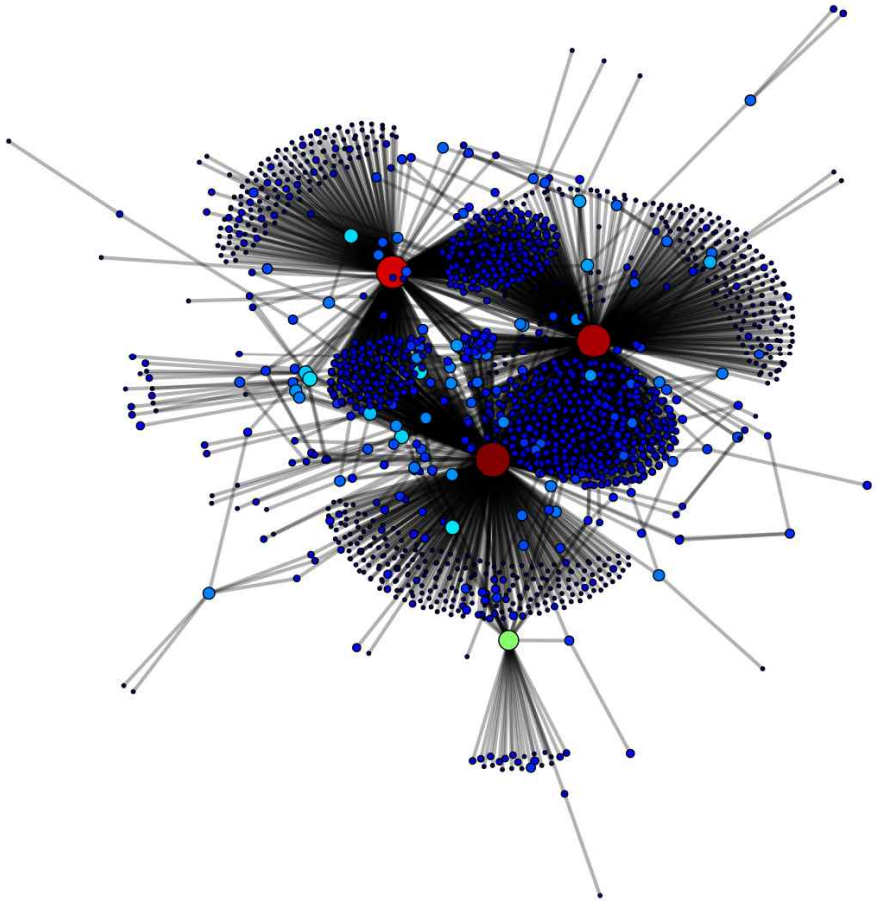
Community size distribution



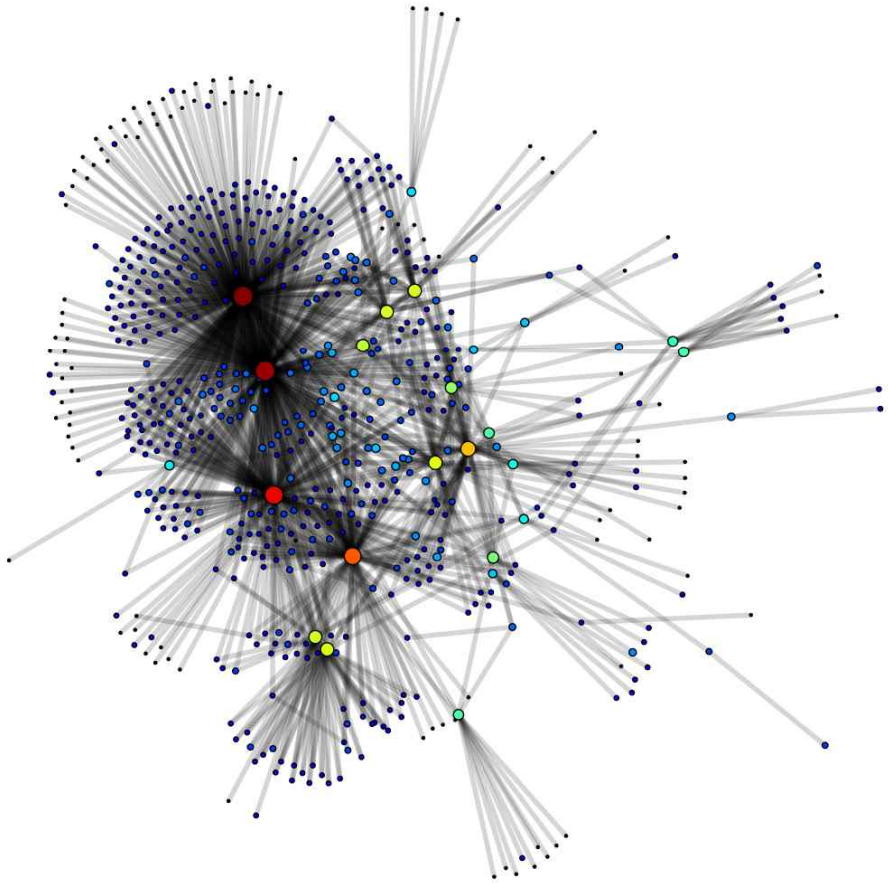
Jaccardi Index



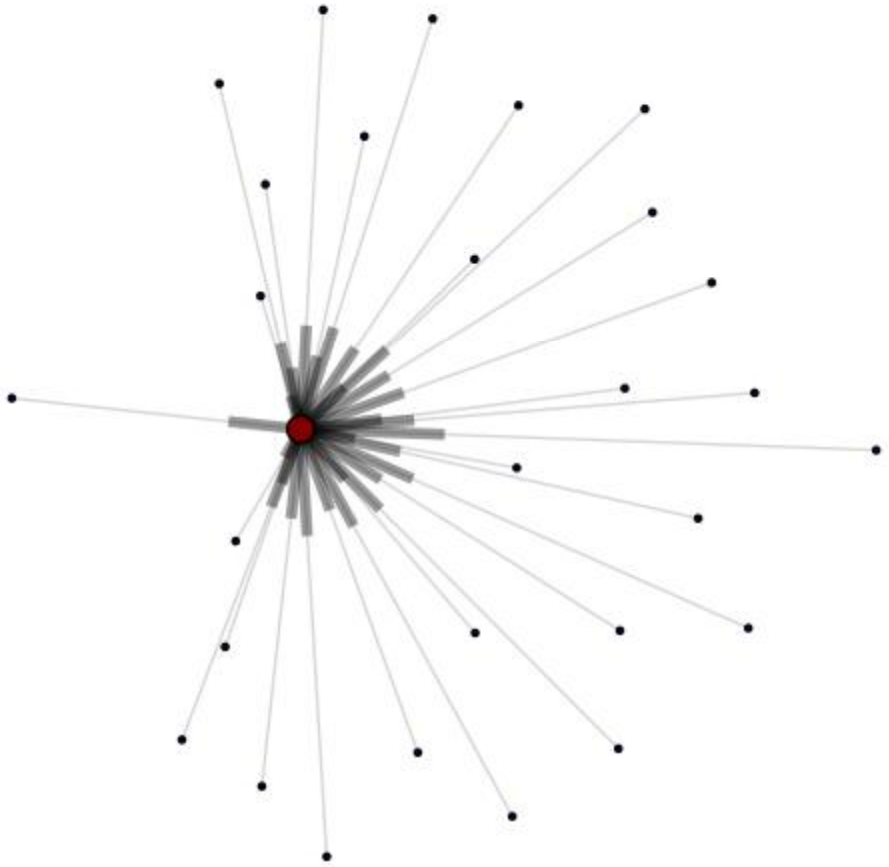
The Largest Community



Korean ISPs



A Star Topology in Ukraine



Summary & Future Work

- Finally, we have a tool for consistent community identification
- Preliminary analysis looks promising
  - AS graph and Cyworld group evolution
- Analytical framework for convergence?
- Overlapping communities?
- Community merging? Splitting?