Measurement and Analysis of Online Social Networks

Presented by Lin Gong

April 2nd, 2015

Roadmap

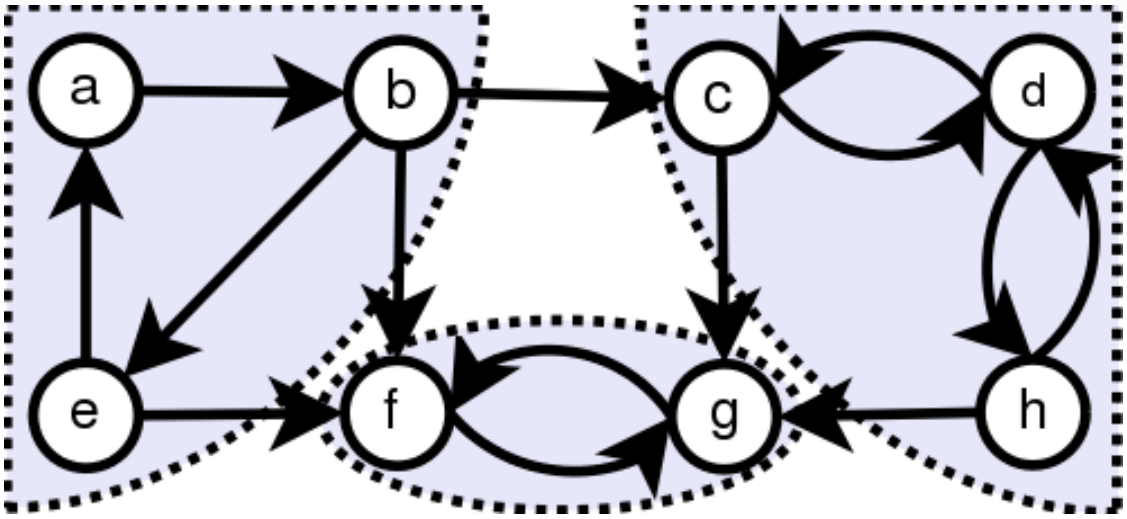
- Introduction
  - Offline social networks
  - Basic knowledge
- Motivation
  - Why?
- Measurement Methodology
  - Collect data
  - Coverage evaluation
- Analysis of Network Structure
  - Power-law node degree
  - Correlation of indegree and outdegree
  - Core and fringe

What are online social networks?

Online social networks

- Graphs of people based on Web
- Online acquaintance
  - Share interests and trust
  - Like contributed content
- Online friends may have never met
- Different ways of sharing

Some basic knowledge



- SCC: strongly connected component
- WCC: weakly connected component
- Indegree: the number of head endpoints adjacent to the node
- outdegree: the number of tail endpoints adjacent to the node

Roadmap

- ~~Introduction~~
  - ~~Offline social networks~~
  - ~~Basic knowledge~~
- Motivation
  - Why?
- Measurement Methodology
  - Collect data
  - Coverage evaluation
- Analysis of Network Structure
  - Power-law node degree
  - Correlation of indegree and outdegree
  - Core and fringe

Why study social networks?

- Look
- Guic
- Explo
- To te

This work

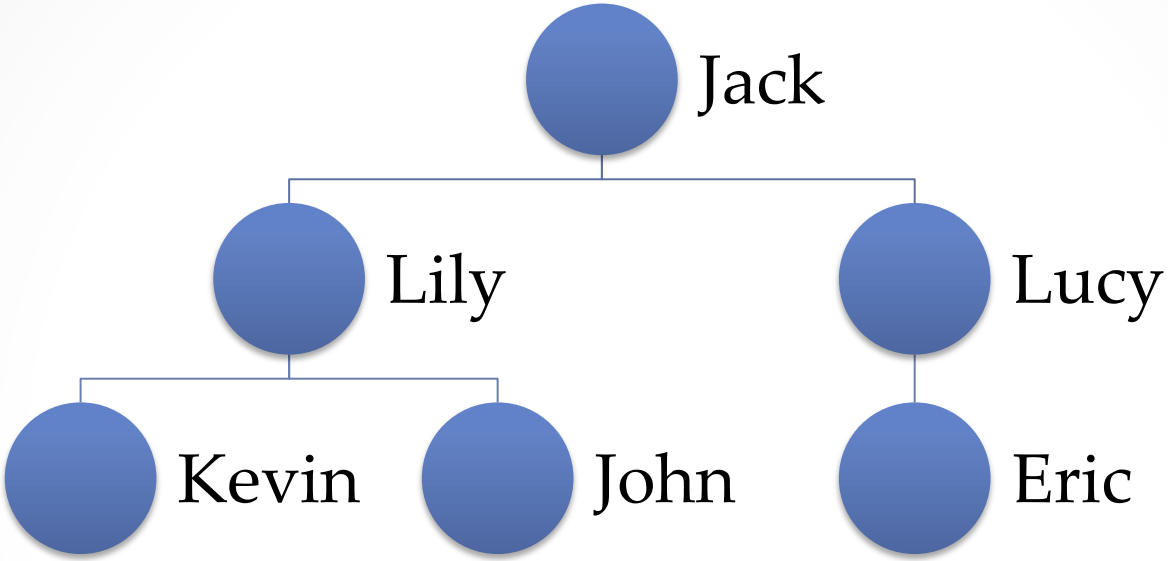
A large-scale measurement study and analysis of the structure of four popular online social networks:

- Flicker – Photo sharing
- YouTube – Video sharing
- LiveJournal – Blogging site
- Orkut – Social networking site

Roadmap

- ~~Introduction~~
  - ~~Offline social networks~~
  - ~~Basic knowledge~~
- ~~Motivation~~
  - ~~Why?~~
- Measurement Methodology
  - Collect data
  - Coverage evaluation
- Analysis of Network Structure
  - Power-law node degree
  - Correlation of indegree and outdegree
  - Core and fringe

How do they collect data?



- Select seed users.
  - Crawl all his/her friends.
  - Add new users to the list.
- Continue until all known users are crawled.
- Perform a BFS of the graph.

Coverage Evaluation: Flickr

- Select random users by guessing their user names which have format:

  ########@N00

- Fraction of connected users is 27%, disconnected is 73%.
- Among the disconnected users:
  - 80% users have fewer than 3 links.
- Cover a large portion of the large WCC.

Coverage Evaluation: LiveJournal

- Use API provided by LiveJournal.
- Use a feature of LiveJournal that returns random users.
- Fraction of disconnected users is only 5%.
- Covers almost the complete population.

Coverage Evaluation: Orkut

- Orkut was fully connected but ended crawling early, so only get a 11.3% subset.
- The representativeness of the dataset:
  - Perform multiple crawls from different seeds.
- Exist sampling bias caused by the partial BFS.

Coverage Evaluation: YouTube

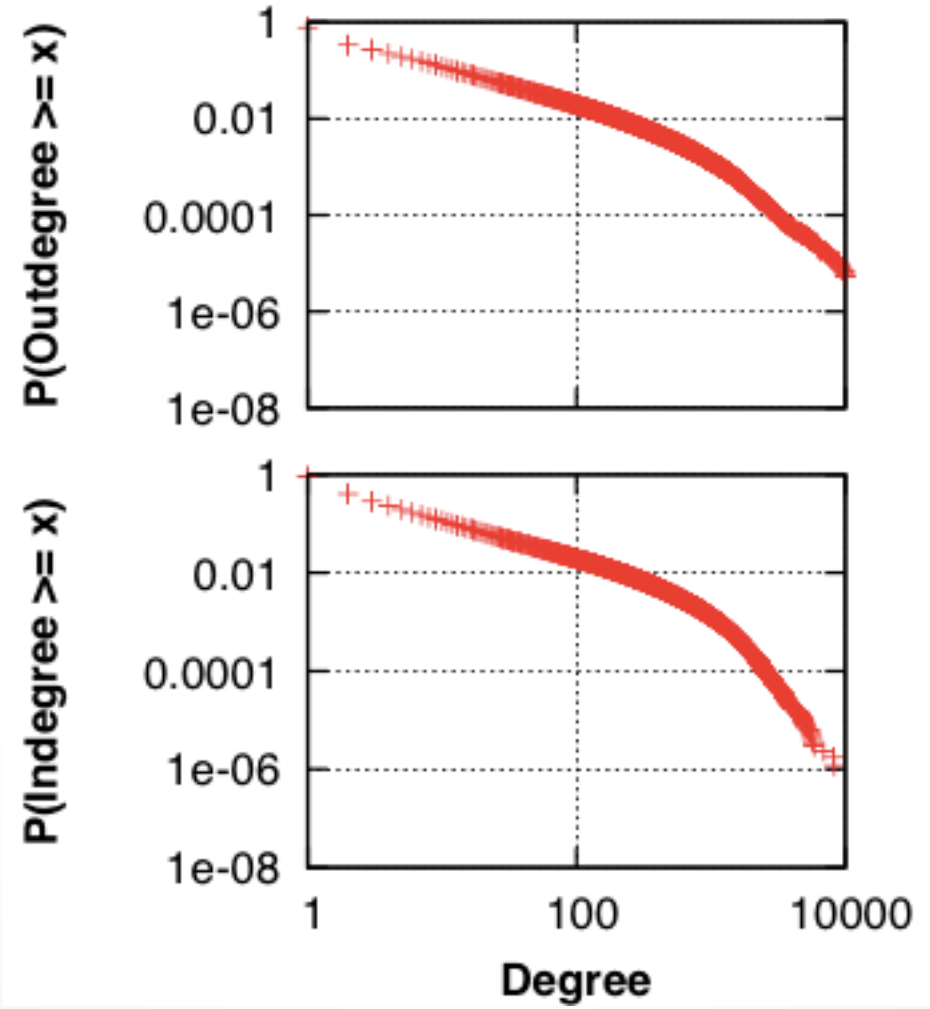
- Unable to estimate the entire YouTube population.
- May not contain some nodes in the large WCC, but the fraction is likely to be small.
- Cover a large portion of the large WCC.

Roadmap

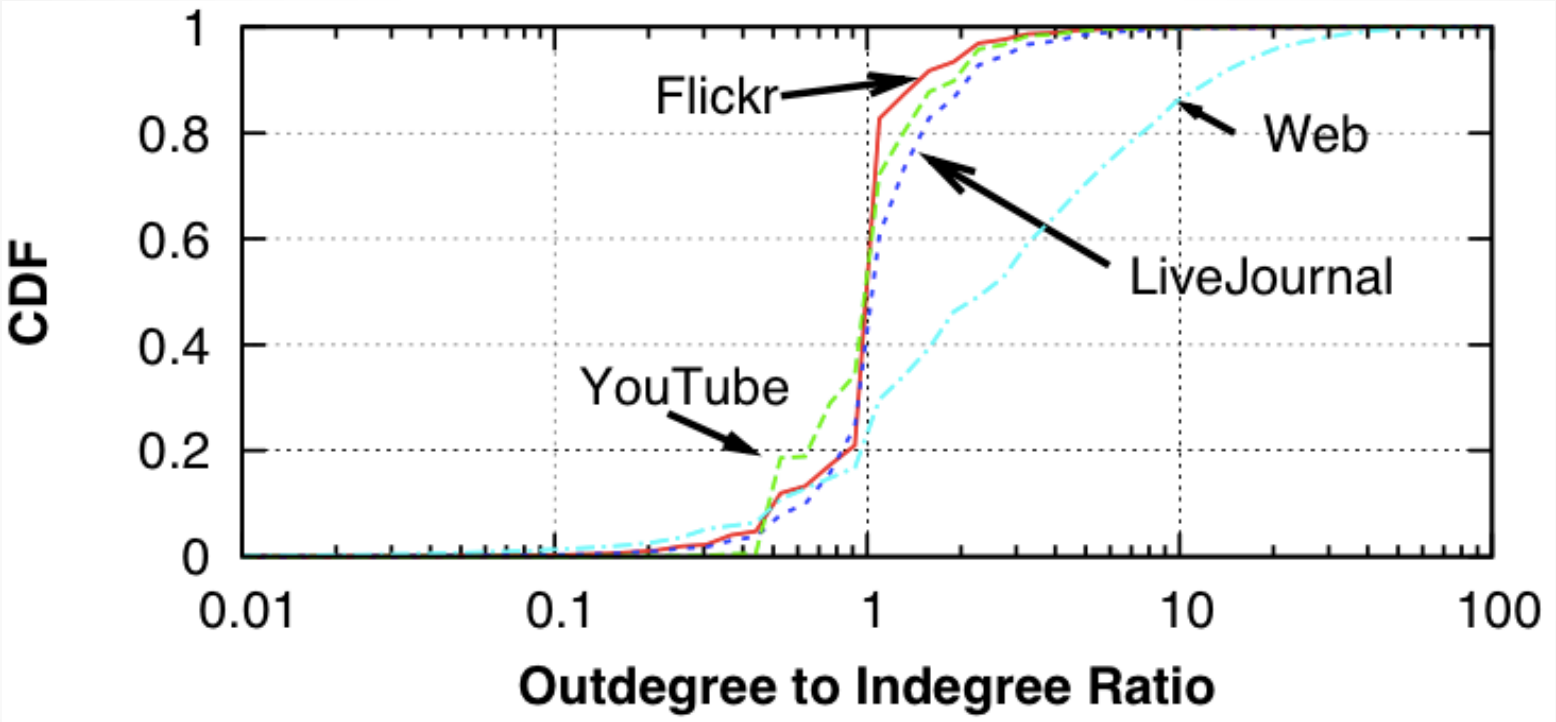
- ~~Introduction~~
  - ~~Offline social networks~~
  - ~~Basic knowledge~~
- ~~Motivation~~
  - ~~Why?~~
- ~~Measurement Methodology~~
  - ~~Collect data~~
  - ~~Coverage evaluation~~
- Analysis of Network Structure
  - Power-law node degree
  - Correlation of indegree and outdegree
  - Core and fringe

Power-law node degrees

Log-log Plot of outdegree and indegree complementary cumulative distribution function(CCDF)



Correlation of indegree&outdegree

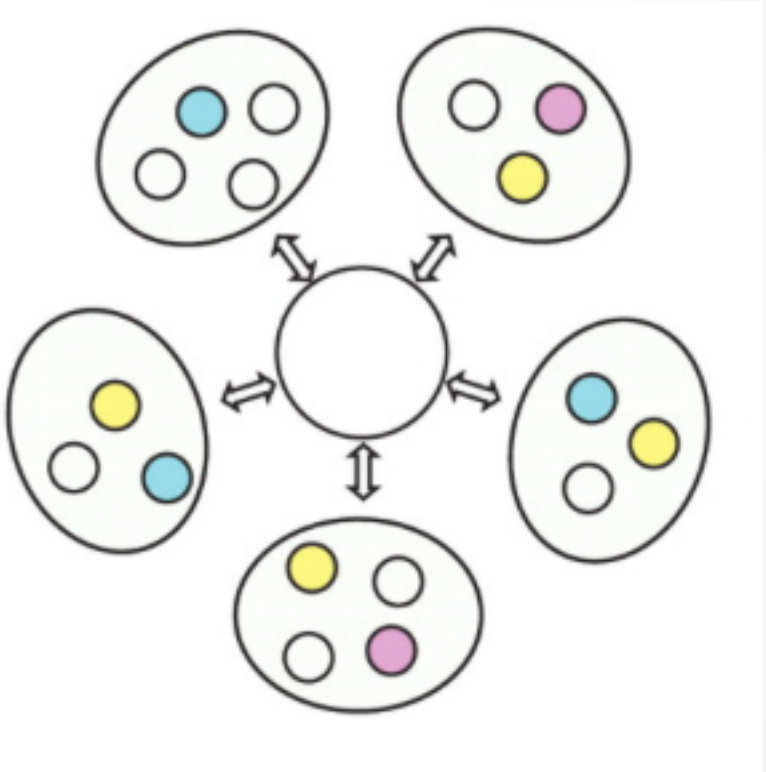


- Social networks:
  - nodes with high outdegrees -> high indegrees.
- The high correlation is caused by the high symmetry.

Densely connected core

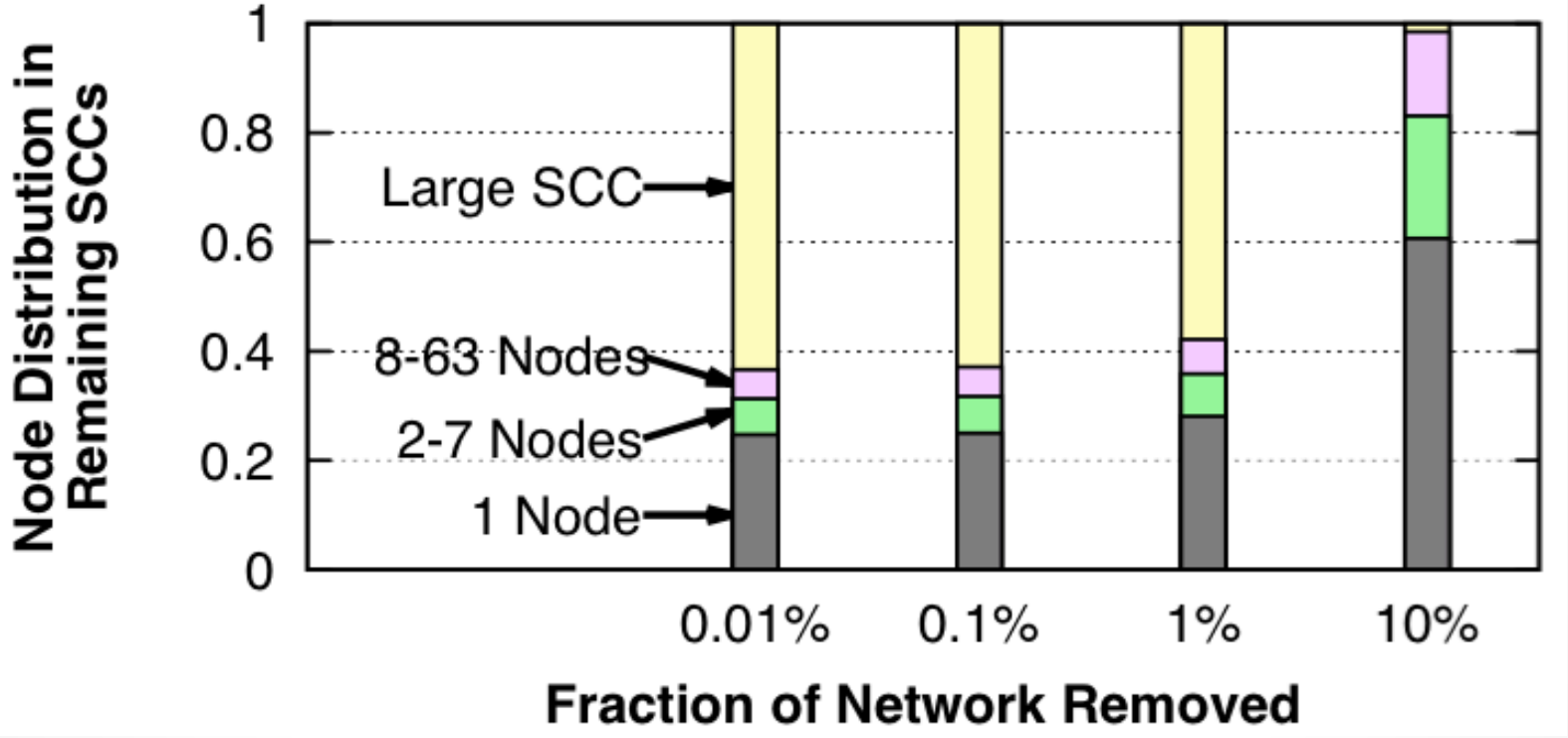
What is a core?

- The core must be necessary for the connectivity of the network.
- The core must be strongly connected with a relatively small diameter.



Densely connected core

- Does core exist?



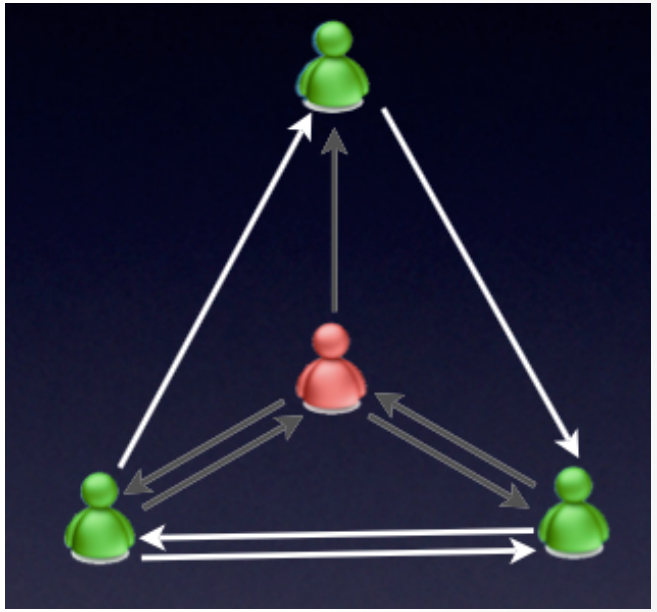
The graphs have a densely connected core comprising of between 1% and 10% of the highest degree nodes.

Tightly clustered fringe

- Clustering coefficient:

$$\frac{\text{Number of links between friends}}{\text{Number of links that could exist}}$$

C= 4/6 = 0.667



Tightly clustered fringe

Network	C	Ratio to Random Graphs	
		Erdös-Rényi	Power-Law
Web [2]	0.081	7.71	-
Flickr	0.313	47,200	25.2
LiveJournal	0.330	119,000	17.8
Orkut	0.171	7,240	5.27
YouTube	0.136	36,900	69.4

- The clustering coefficient of social networks
  - 10,000 times more clustered than random graphs.
  - 5-50 times more than random power-law graph.

Summary

- The first large-scale study of multiple online social networks.
- Perform BFS to collect data.
- Analysis of the structure of social networks:
  - The degree distribution conforms to power-law.
  - High correlation between indegree and outdegree.
  - Exist densely connected core and tightly clustered fringe.

Thanks!

Q & A

References

- http://www.mpi-sws.org/~mmarcon/SocialNetworks-IMC.pdf
- http://socialnetworks.mpi-sws.org
- http://www.ccs.neu.edu/home/amislove/slides/SocialNetworks-IMC-slides.pdf