Growth of the Flickr Social Network

Alan Mislove[†‡] Hema Swetha Koppula[¶] Krishna Gummadi[†]

Peter Druschel[†] Bobby Bhattacharjee[§]

[†]MPI-SWS
[‡]Rice University

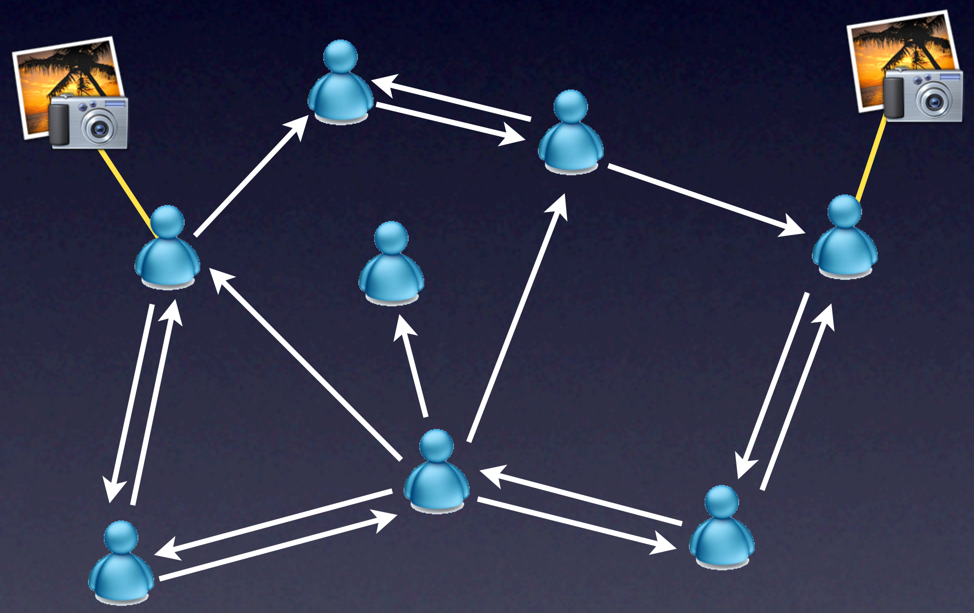
[¶]IIT Kharagpur
[§]University of Maryland

WOSN 2008

Online social networks

- Popular way to connect, share content
    - Among most visited sites on Web
    - Users: Orkut (60 M), LiveJournal (5 M)

- Unique opportunity to dynamics of large, complex social networks



Why study social network growth?

- Online social networks share many structural properties
  - Significant clustering, small diameter, power-law degrees
  - Similar underlying growth processes?

- Proper understanding of growth can
  - Provide insights into structure
  - Predict future growth
  - Model arbitrary-sized networks

- Most work to-date relies on theoretical models
  - Not known if they predict actual growth

This work

- Use a measurement-driven approach to understand growth
- Present large-scale measurement of Flickr network growth
  - ~1 M new users, ~10 M new links
- Look for underlying cause of structural characteristics
  - High symmetry
  - Power-law node degree
  - Significant local clustering

Contributions

- Methodology to collect large-scale network growth data
  - Measured both Flickr and YouTube
- Make data available to researchers
  - Much larger scale, higher granularity than existing data sets
  - Already in use
- Initial analysis
  - Examine high-level properties of growth data
  - Test whether data is consistent with existing models

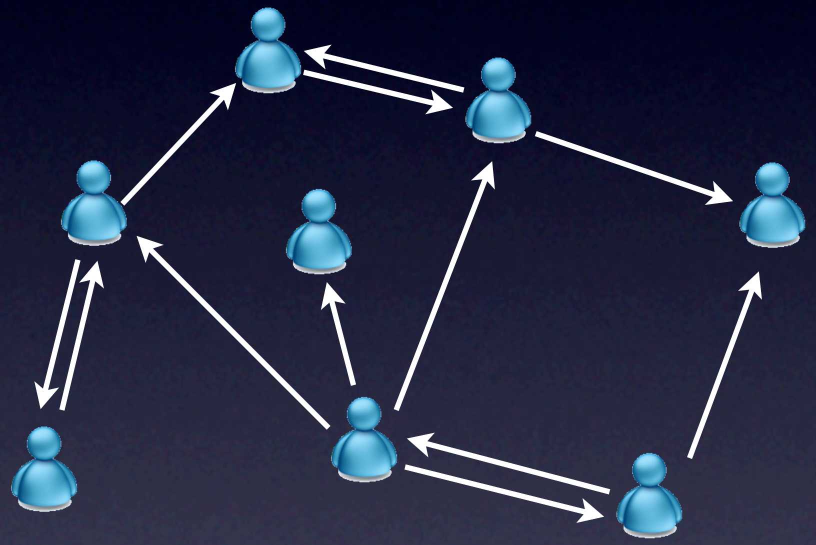
Rest of the talk

- Measuring social network growth
- Analyzing growth properties
- Related work

Crawling social networks

- Flickr reluctant to give out data
  - Cannot enumerate user list
  - Instead, performed crawls of user graph
- Picked known seed user
  - Crawled all of his friends
  - Added new users to list
- Continued until all reachable users crawled
- Effectively performed a BFS of graph

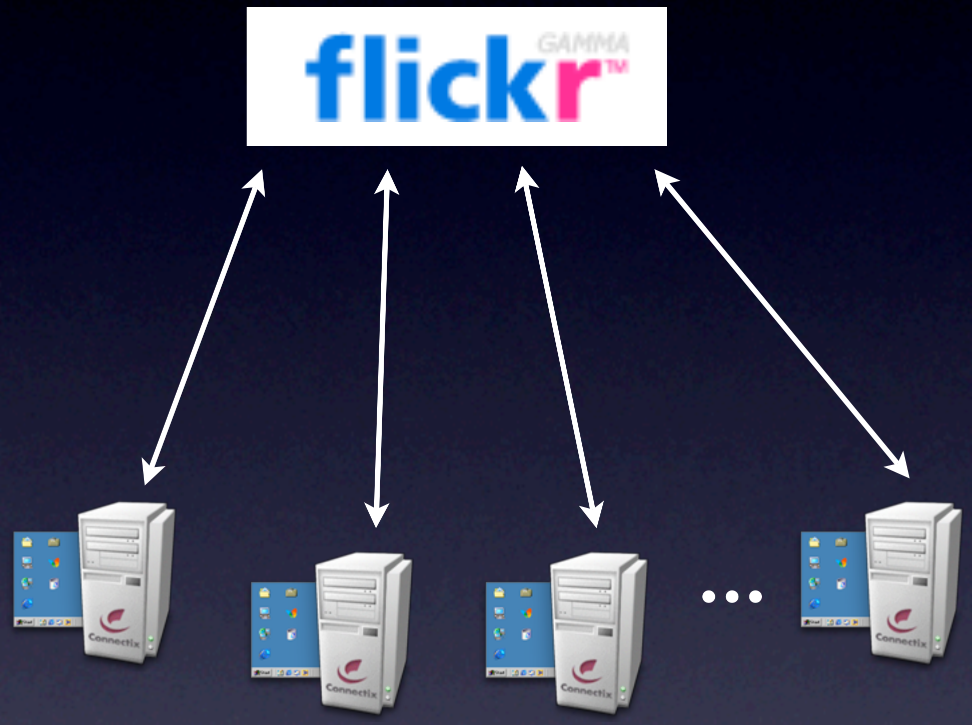
Crawling social networks



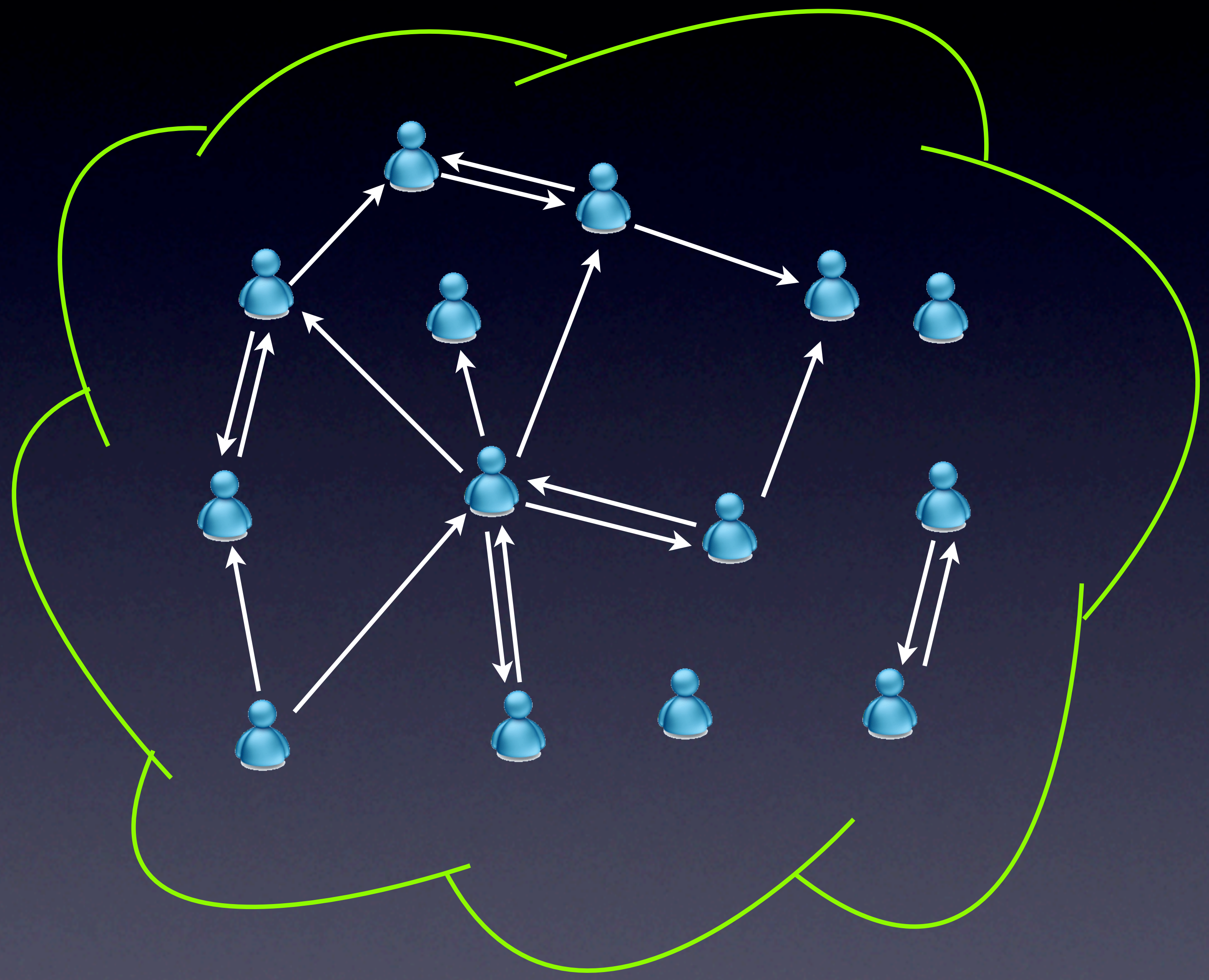
- Flickr reluctant to give out data
  - Cannot enumerate user list
  - Instead, performed crawls of user graph
- Picked known seed user
  - Crawled all of his friends
  - Added new users to list
- Continued until all reachable users crawled
- Effectively performed a BFS of graph

Observing growth

- Crawls subject to rate-limiting
  - Discovered appropriate rate
- Crawled using cluster of 58 machines
  - Using Flickr API
- Result: could complete crawl in 1 day
- Repeated daily for 3 months
  - Revisited all previously discovered users
  - Looked for new links, users

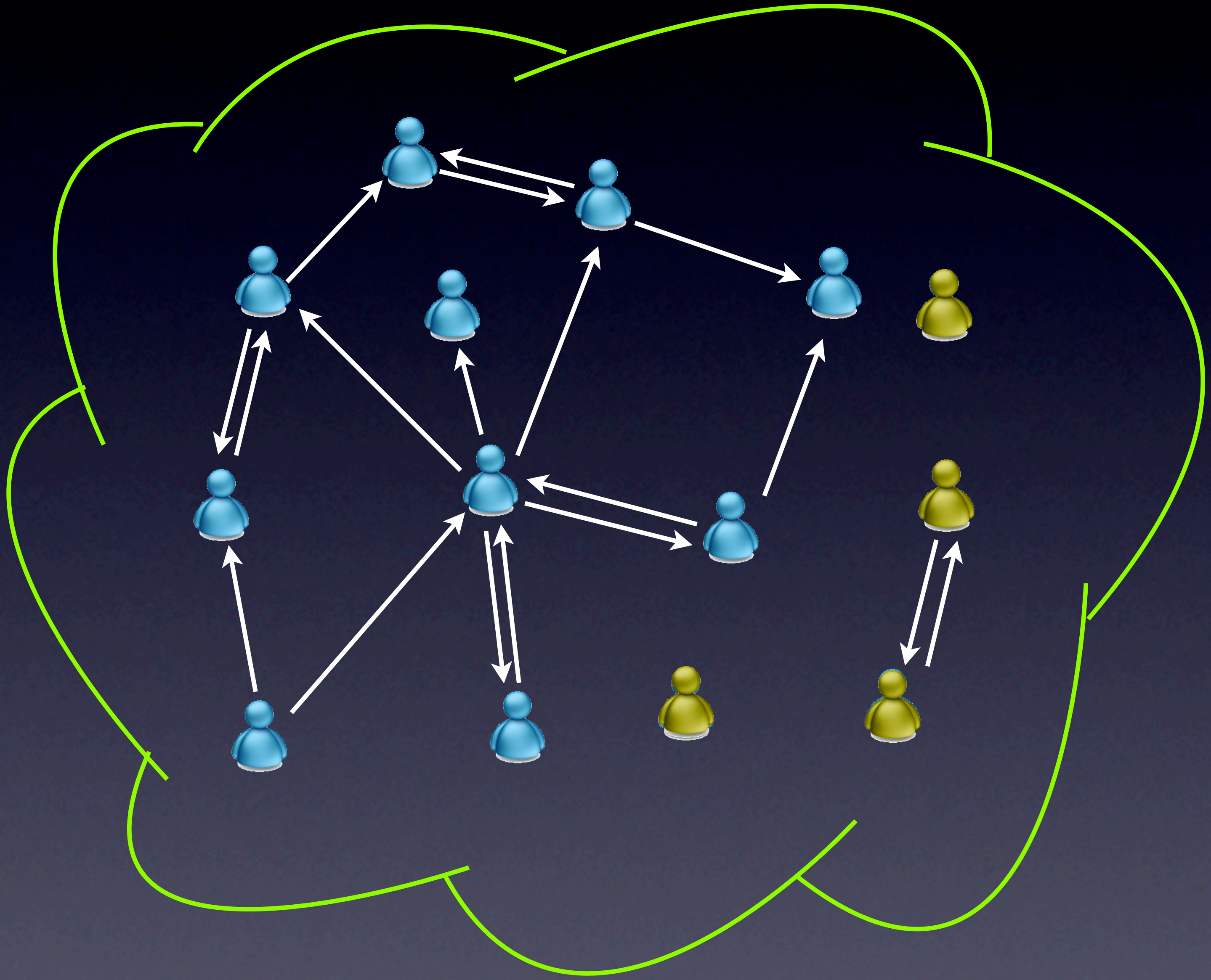


How much were we able to crawl?



- Users don't necessarily form single WCC
  - Disconnected users
- Estimate coverage by selecting random users
  - Result: 27% coverage
- But, disconnected users have very low degree
  - 90% have no outgoing links

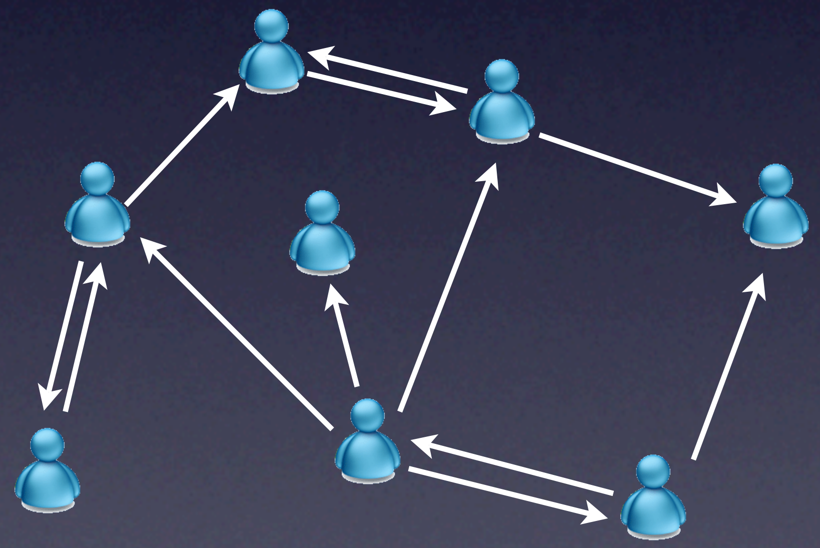
How much were we able to crawl?



- Users don’t necessarily form single WCC
  - Disconnected users
- Estimate coverage by selecting random users
  - Result: 27% coverage
- But, disconnected users have very low degree
  - 90% have no outgoing links

Limitations to growth data

- Newly discovered users may have existing links
  - Don't know when existing links were created
  - Only count links we observed being created

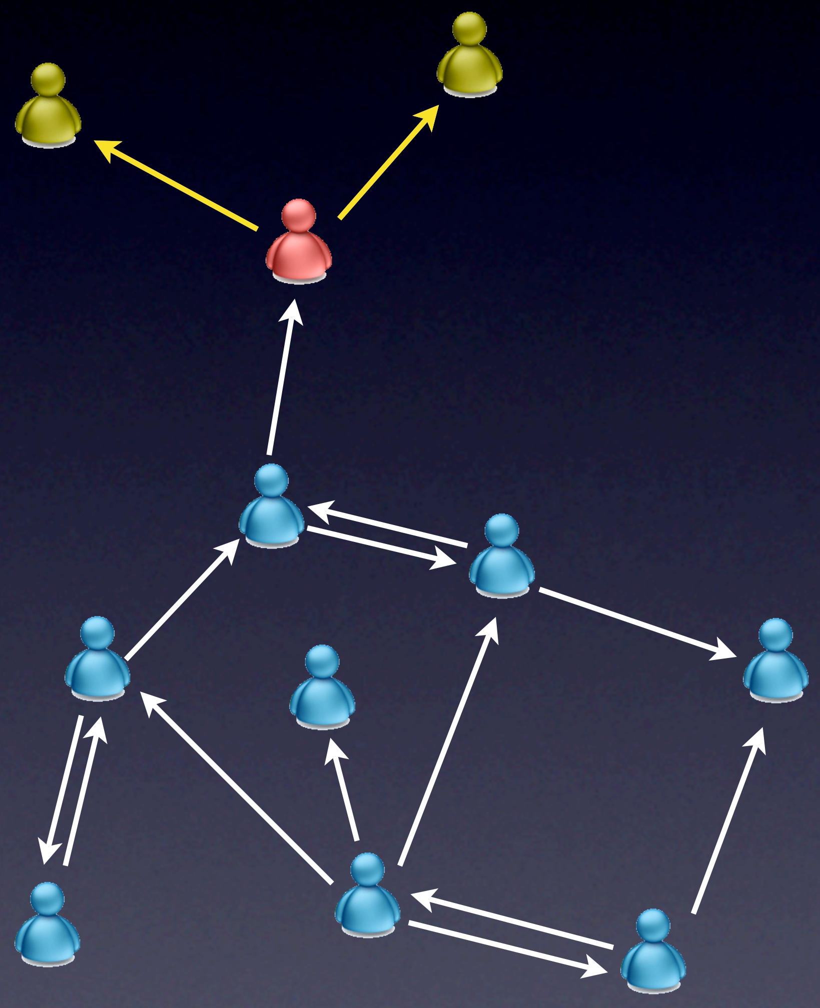


- Crawls have resolution of 1 day
  - Can't tell order of link creation within a day

Limitations to growth data

- Newly discovered users may have existing links
  - Don’t know when existing links were created
  - Only count links we observed being created

- Crawls have resolution of 1 day
  - Can’t tell order of link creation within a day



Rest of the talk

- ~~Measuring social network growth~~
- Analyzing growth properties
- Related work

Growth data characteristics

- Crawled Flickr daily for over 3 months
  - Nov. 2 - Dec. 3, 2006 and Feb. 3 - May 18, 2007
- Observed ~1 M new users and ~10 M new links
  - Network grew from 17 M to 33 M links
  - Growth rate of 455% per year
- Link addition dominates removal
  - 2.43:1 ratio (conservative)
  - Focus only on link addition

Network growth questions

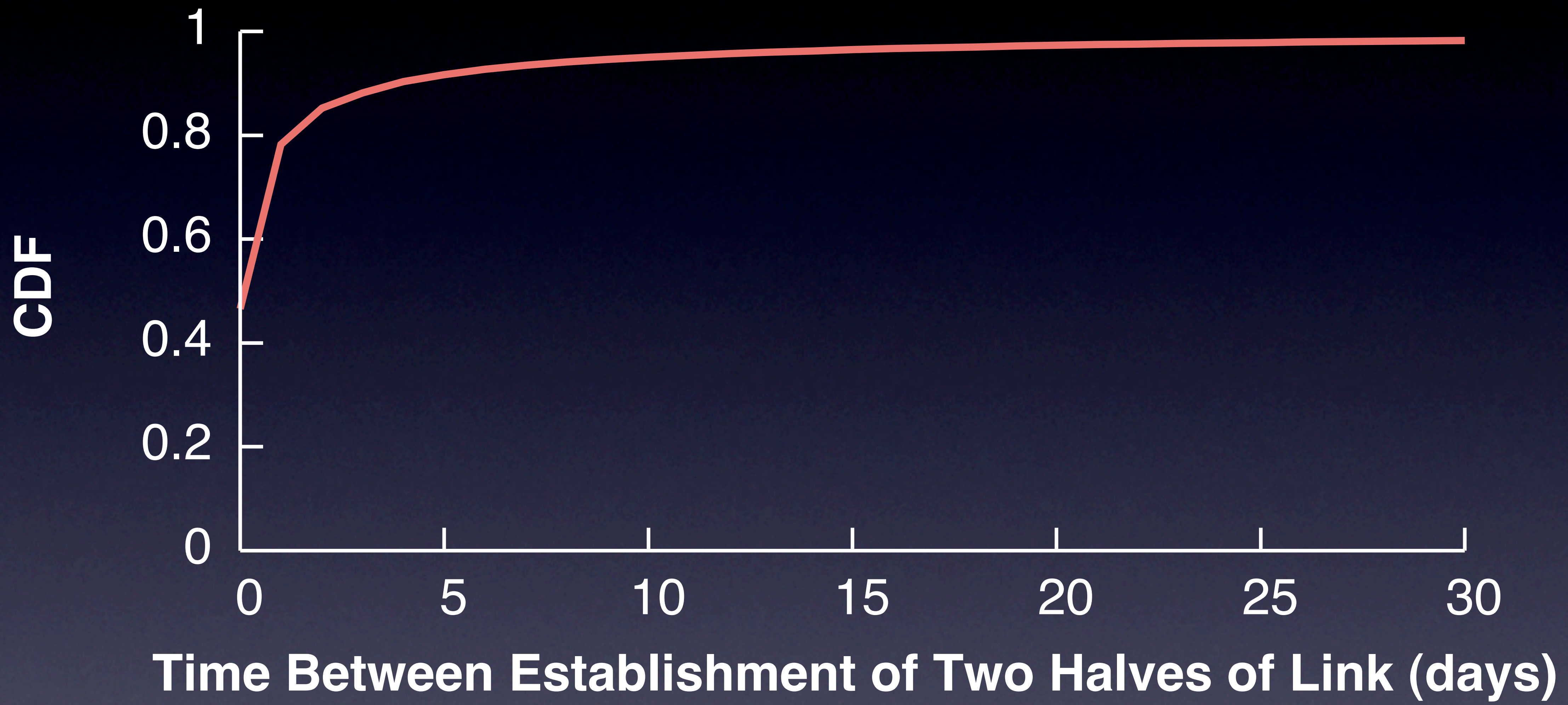
- How does growth lead to observed structural properties?
- Is growth consistent with a known model?

- Networks have high symmetry
  - What causes symmetric links to form?

- Networks follow power-laws
  - Which users create and receive new links?
  - Does it happen via *preferential attachment*?

- Networks have significant local clustering
  - Much higher than random power-law graphs
  - How do users select new destinations?

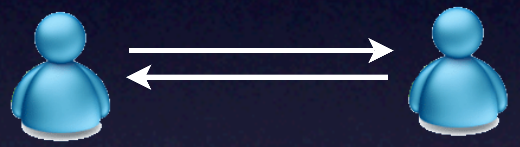
How quickly do symmetric links form?



- Over 80% of symmetric links created within 48 hours

Reciprocity

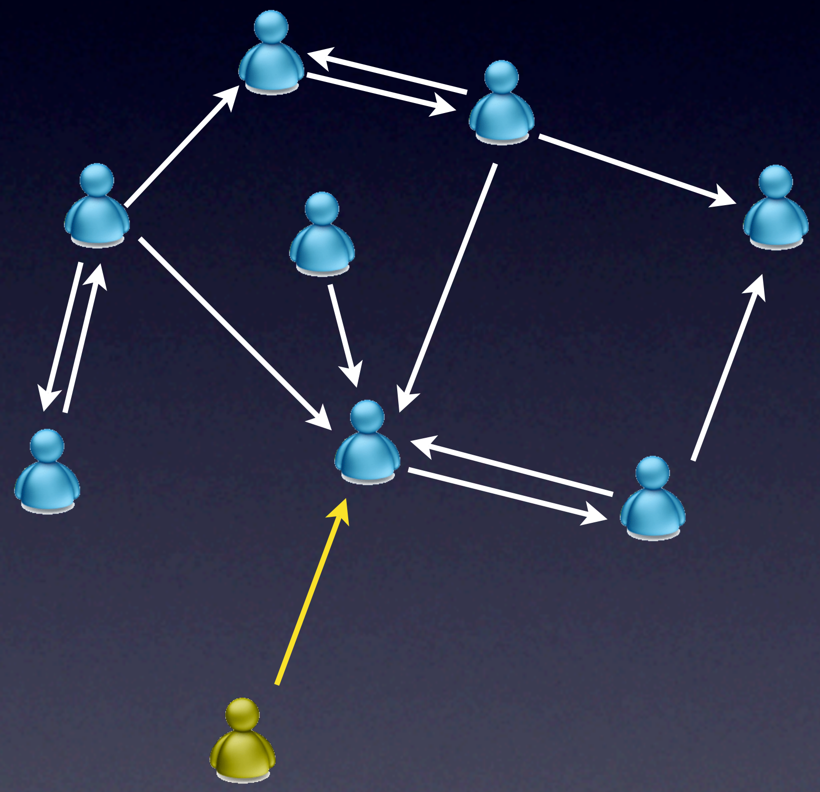
- Users can create link in response to incoming link
  - “Out of courtesy”
  - Known in sociology



- Flickr emails users about new incoming links

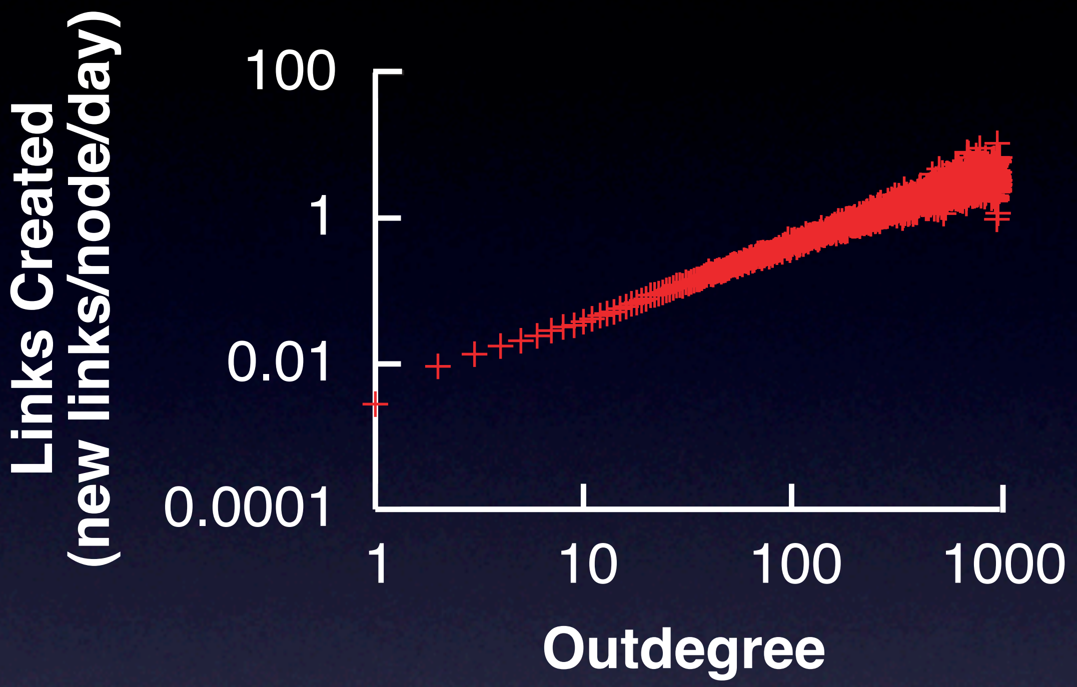
- Data consistent with *reciprocity* causing high level of link symmetry

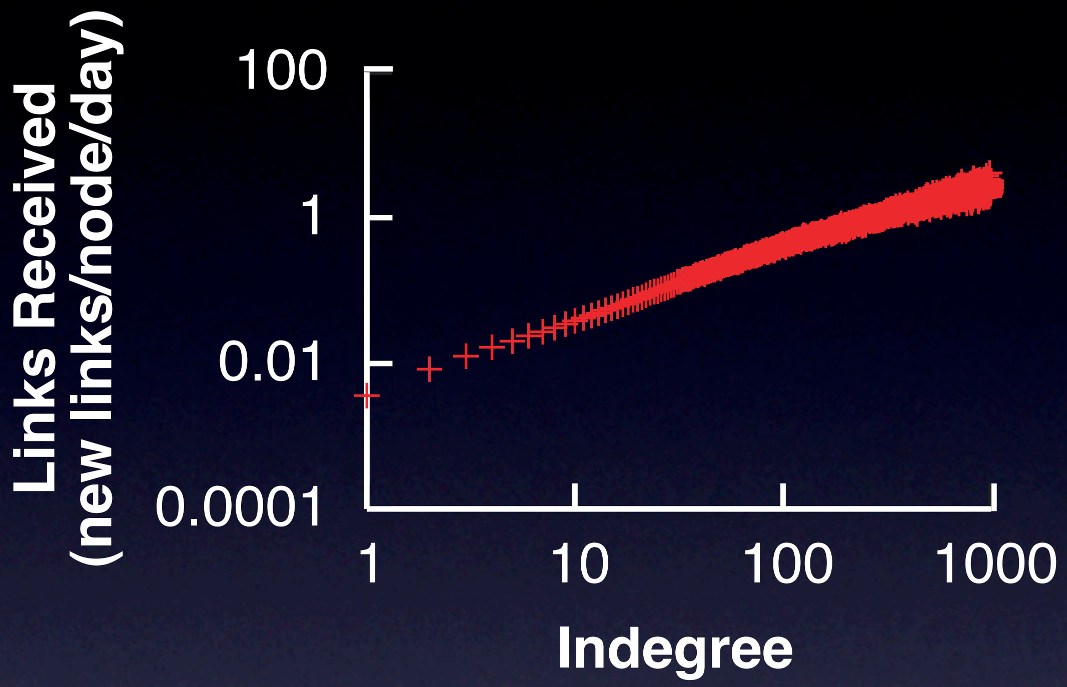
Preferential attachment



- Model for creating power-law networks
  - Known as "cumulative advantage" or "rich get richer"
- New links go *preferentially* to nodes with many links
- For directed networks, we define
  - *Preferential creation*
  - *Preferential reception*

Is preferential attachment happening?





- Yes, linear correlation between
  - Links created and outdegree (preferential creation)
  - Links received and indegree (preferential reception)

- Is this consistent with a known model?
  - Both global and local models have been proposed

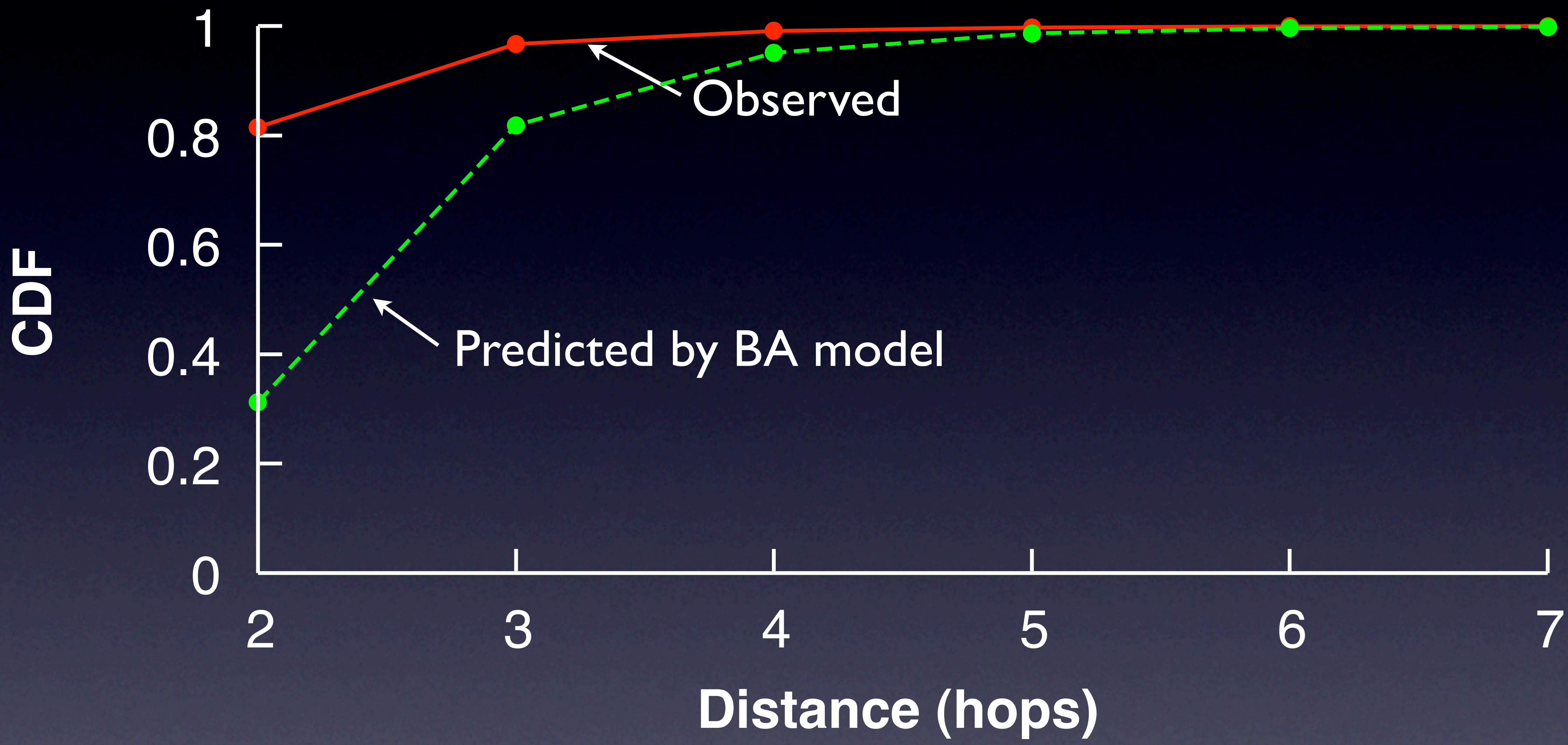
Barabasi-Albert (BA) model

- Well-known model for creating power-law networks
- Uses global preferential attachment
  - Destination selected using global weighted ranking

$$P(x) = \frac{d_x}{\Sigma\ d_i}$$

- Is data consistent with such a global process?
  - Look for evidence using distance between source and destination

Does proximity matter?



- New friends much closer than BA model predicts
    - Models which take into account local rules may be more accurate

Implications of network growth

- Observed growth of a large, complex social network
- Found multiple growth processes at work
  - Reciprocity leads to high symmetry
  - Preferential attachment leads to power-law degrees
  - Proximity bias leads to local clustering
- But, data inconsistent with global BA model
- Future work: Modeling complex network growth
  - Based on local rules
  - Verify consistency of data with other proposed models

Related work

- Growth models
  - Preferential attachment [Science'99]
  - Random walks [Phya.A'04]
  - Common neighbors [Phys.Rev.E'01]
- Small-scale empirical studies
  - Scientific collaboration networks [Phys.Rev.E'01,Euro.Phy.Ltrs'04]
  - Email networks [Science'06]
  - Movie actor networks [J.Stat.Mech.'06]

Summary

- Presented first large-scale study of online social network growth
- Collected data covering ~1 M new users, ~10 M new links
- Found high-level growth processes at play
  - Growth via local, rather than global, processes
- Data sets are available to researchers
  - Many already using data (72 researchers, including sociologists!)
  - Also have growth data for YouTube network

Questions?

Data sets available from:

http://socialnetworks.mpi-sws.org