# Structure and Growth of Online Social Networks

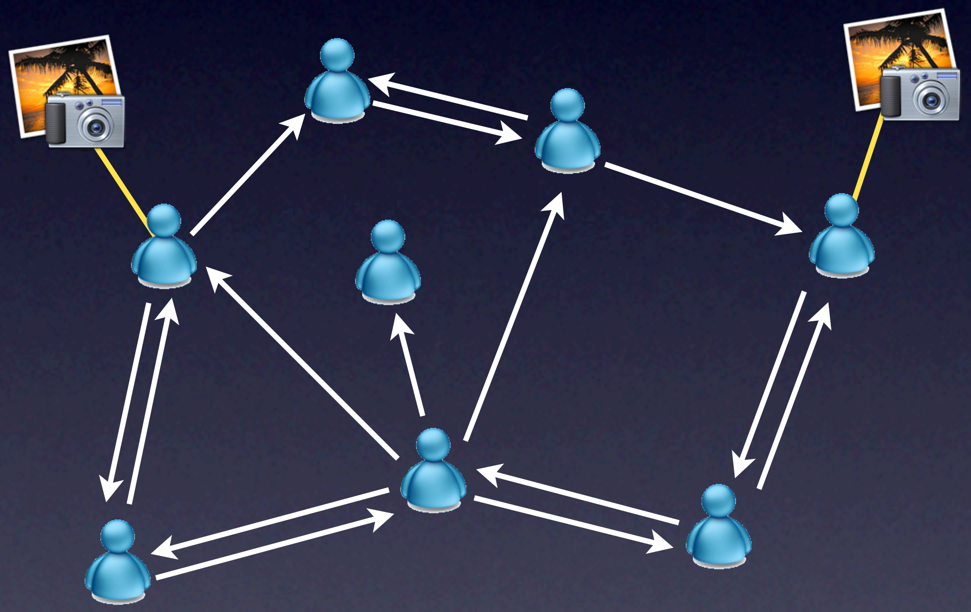
Alan Mislove[†‡] Krishna Gummadi[†] Peter Druschel[†]

[†]Max Planck Institute for Software Systems
[‡]Rice University

MSRC Social Networks Workshop
07.12.2007

# Why are social networks interesting?

- Popular way to connect, share content
  - Photos (Flickr), videos (YouTube), blogs (LiveJournal), profiles (Orkut)
  - Orkut (60 M), LiveJournal (5 M)



- Content organized with user-user links
  - Akin to Web's page-page links
  - Social network *structure* influences how content is shared

# Our research agenda

- Observe and understand online social networks
  - Measure static structural properties
  - Observe network growth
  - Characterize information flow

- Leverage social networks to build better systems
  - Trust can be used to solve security problems
  - Shared interest can improve content location

# Our research agenda

- Observe and understand online social networks
  - Measure static structural properties
  - Observe network growth
  - Characterize information flow

- Leverage social networks to build better systems
  - Trust can be used to solve security problems
  - Shared interest can improve content location

# Computational sociology

- Marrying network measurement with sociology
  - Able to collect data at massive scale
  - Bringing measurement techniques to bear on social networks
- Data we have collected:
  - Structural information on 11.3M users and 328M links
  - Observed over 2.9M new users join and 24M new links created
  - Data on over 5M photos and videos
- All data is (or will be) publicly available

**`http://socialnetworks.mpi-sws.org`**

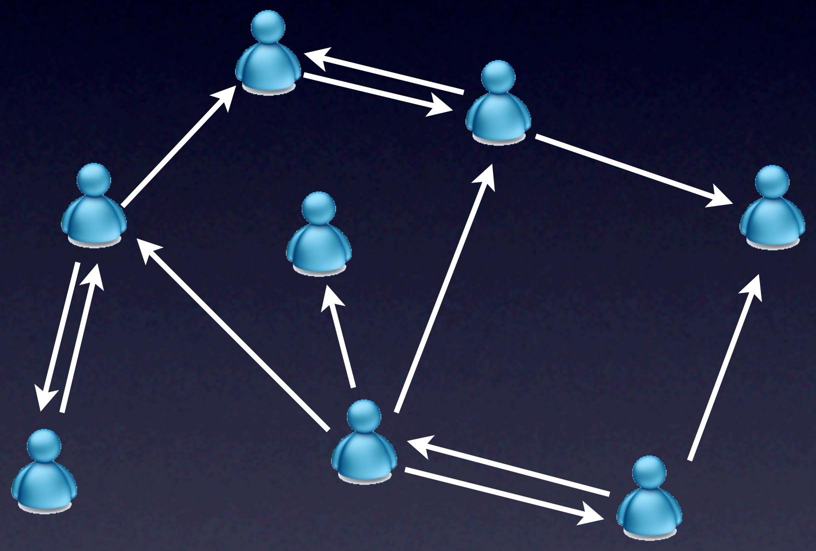
# Part I:

# Analyzing network structure

# Measuring online social networks

- Sites 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 known users crawled
  - Effectively performed a BFS of graph
- Challenging to estimate coverage

# Measuring online social networks



- Sites 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 known users crawled
  - Effectively performed a BFS of graph
- Challenging to estimate coverage

# High-level data characteristics

| | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|
| Number of Users | | | | |
| Avg. Friends per User | | | | |

- Able to crawl large portion of networks
- Node degrees vary by orders of magnitude
  - However, networks share many key properties

# High-level data characteristics

| | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|
| Number of Users | 1.8 M | 5.2 M | 3.0 M | 1.1 M |
| Avg. Friends per User | | | | |

- Able to crawl large portion of networks
- Node degrees vary by orders of magnitude
  - However, networks share many key properties

# High-level data characteristics

| | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|
| Number of Users | 1.8 M | 5.2 M | 3.0 M | 1.1 M |
| Avg. Friends per User | 12.2 | 16.9 | 106.1 | 4.2 |

- Able to crawl large portion of networks
- Node degrees vary by orders of magnitude
    - However, networks share many key properties

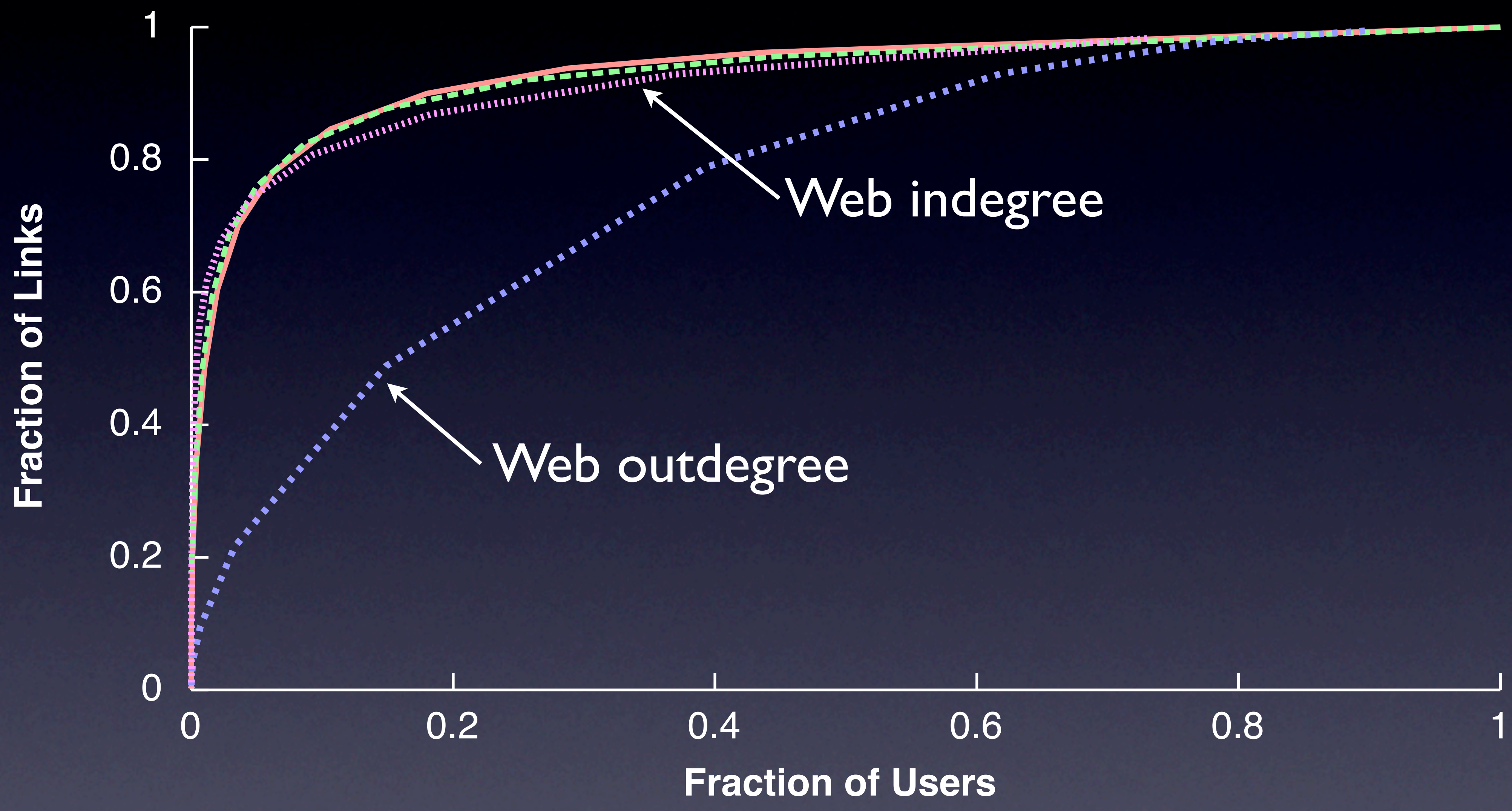
# Are online social networks power-law?

|  | Outdegree γ | Indegree γ |
|---|---|---|
| Web [INFOCOMM'99] | 2.09 | 2.67 |
| Flickr | 1.74 | 1.78 |
| LiveJournal | 1.59 | 1.65 |
| Orkut | 1.50 | 1.50 |
| YouTube | 1.63 | 1.99 |

- Estimated coefficients with maximum likelihood testing
  - Flickr, LiveJournal, YouTube have good K-S goodness-of-fit

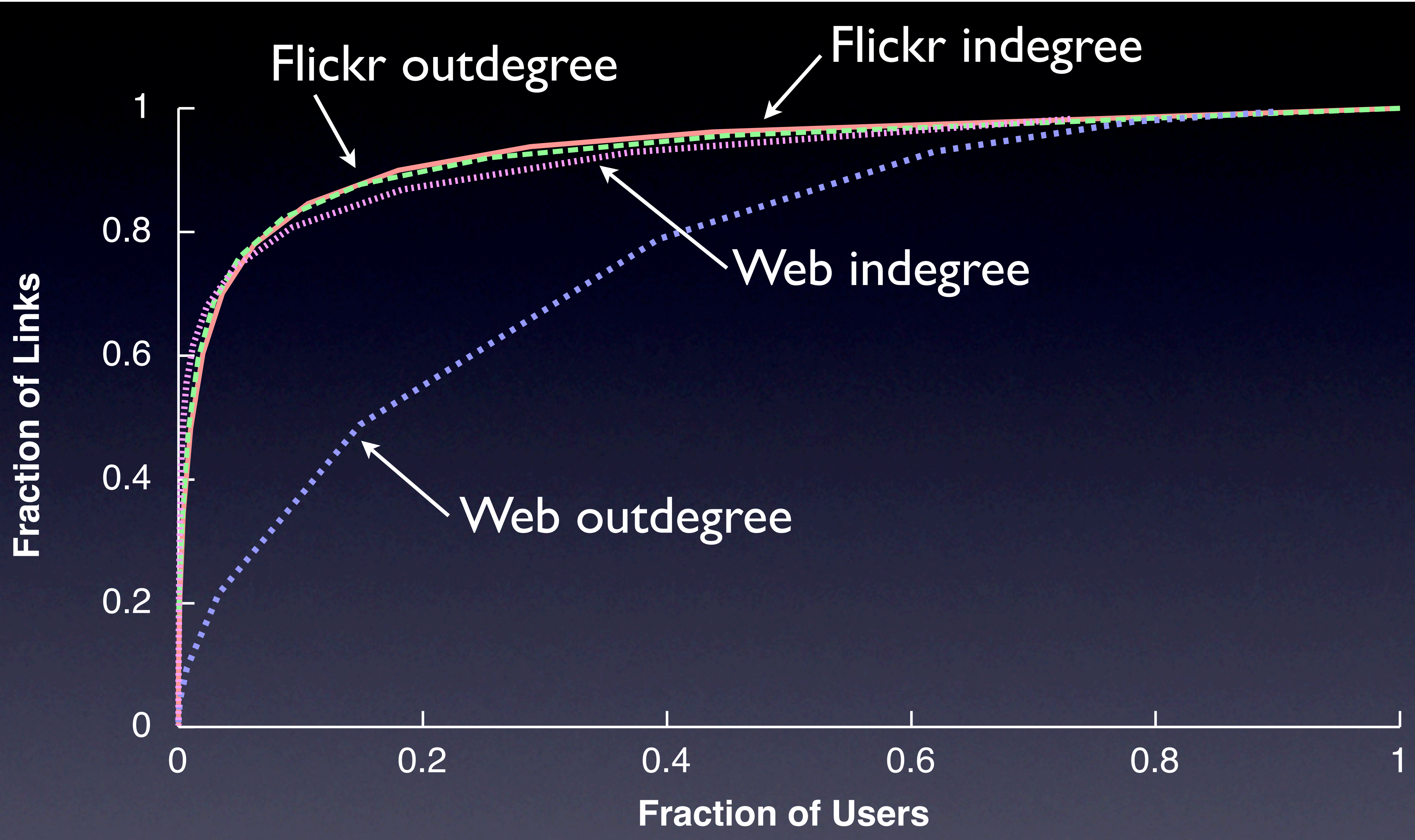
- Similar coefficients imply a similar distribution of in/outdegree
  - Unlike Web

# How are the links distributed?



- Distribution of indegree and outdegree is similar
    - Underlying cause is significant *link symmetry*

# How are the links distributed?



- Distribution of indegree and outdegree is similar
  - Underlying cause is significant *link symmetry*

# Link symmetry

- Social networks show high level of link symmetry
  - Links in most networks are directed

| | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|
| Symmetric Links | | | | |

- High symmetry increases network connectivity
  - Reduces network diameter

# Link symmetry

- Social networks show high level of link symmetry
  - Links in most networks are directed

| | Flickr | LiveJournal | Orkut | YouTube |
|---|---|---|---|---|
| Symmetric Links | 62% | 73% | 100% | 79% |

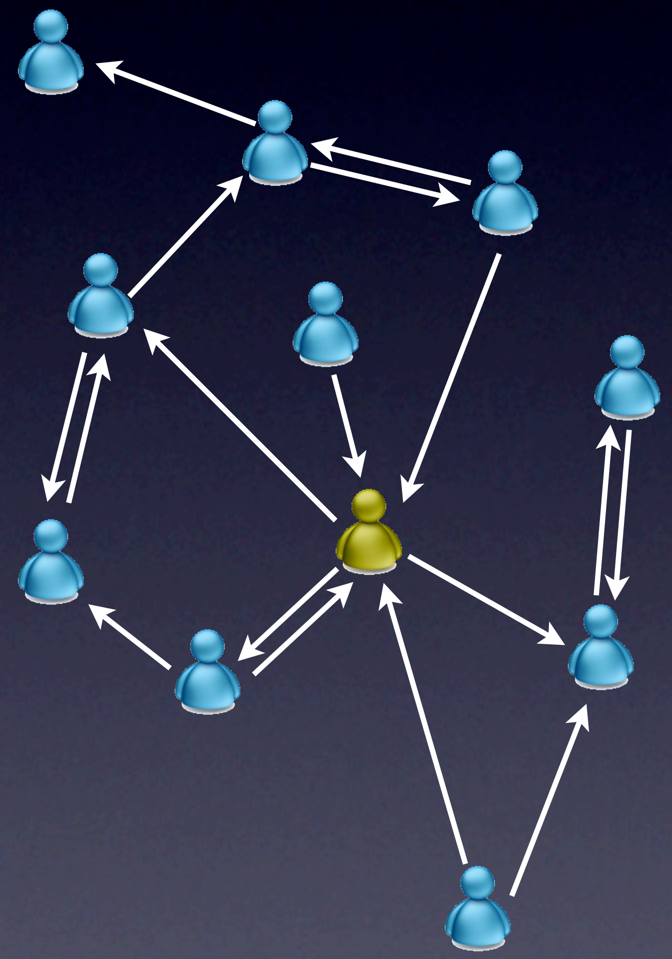
- High symmetry increases network connectivity
  - Reduces network diameter

# Implications of high symmetry

- High link symmetry implies indegree equals outdegree
  - Users tend to receive as many links as the give
- Unlike other complex networks, such as the Web
  - Sites like **`cnn.com`** receive many more links than they give
- Implications is that 'hubs' become 'authorities'
  - May impact search algorithms (PageRank, HITS)
- So far, observed networks are power-law with high symmetry
  - Take a closer look next

# Complex network structure

- What is the high-level structure of online social networks?
  - A jellyfish, like the Internet? [JCN'06]
  - A bowtie, like the Web? [WWW'00]
- In particular, is there a *core* of the network?
  - Core is a (minimal) connected component
  - Removing core disconnects remaining nodes
- Approximate core detection by removing high-degree nodes

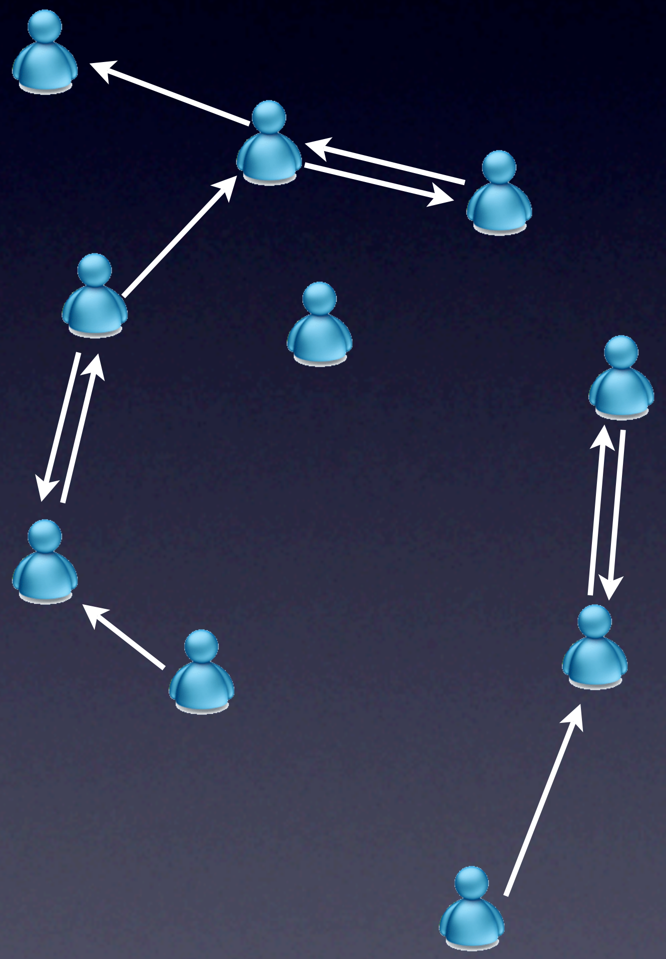


# Complex network structure

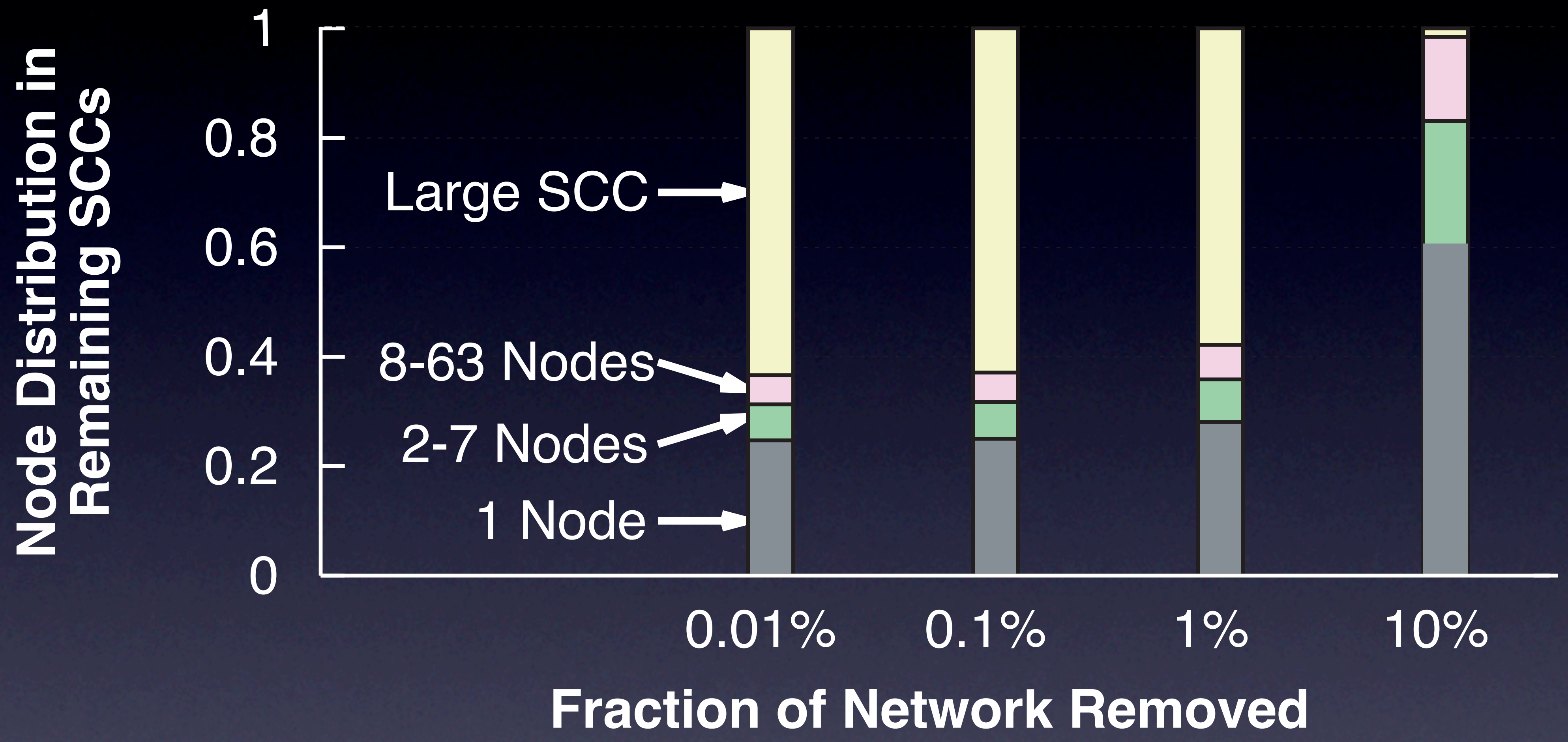
- What is the high-level structure of online social networks?
  - A jellyfish, like the Internet? [JCN'06]
  - A bowtie, like the Web? [WWW'00]

- In particular, is there a *core* of the network?
  - Core is a (minimal) connected component
  - Removing core disconnects remaining nodes

- Approximate core detection by removing high-degree nodes



# Does a core exist?



- Yes, networks contain core consisting of 1-10% of nodes
  - Removing core disconnects other nodes

# Implications of network structure

- Network contains dense core of users
  - Core necessary for connectivity of 90% of users
  - Most short paths pass through core
  - Could be used for quickly disseminating information

- Remaining nodes (fringe) are highly clustered
  - Users with few friends form mini-cliques
  - Similar to previously observed offline behavior
  - Could be leveraged for sharing information of local interest

# Part II:

# Characterizing network growth

# Observing network growth

- Online social networks growing at rapid pace
  - Not possible with Web, Internet

- Offers unique opportunity to observe growth
  - Validate or invalidate existing models
  - Predict future growth

- Also examined evolution of other complex information networks
  - Internet topology: CAIDA archives
  - Wikipedia: Wikimedia archives

# Growth data characteristics

| | Observation Period | Node Growth Rate | Link Growth Rate |
|---|---|---|---|
| Flickr | | | |
| YouTube | | | |
| Wikipedia | | | |
| Internet | | | |

- Crawled social networks repeatedly for months
  - Observed 1.2M new users and 16.8M new links

- Question: What processes are driving network growth?

# Growth data characteristics

| | Observation Period | Node Growth Rate | Link Growth Rate |
|---|---|---|---|
| Flickr | 104 days | 242% | 455% |
| YouTube | 36 days | 145% | 215% |
| Wikipedia | 825 days | 54% | 120% |
| Internet | 1,281 days | 31% | 43% |

- Crawled social networks repeatedly for months
  - Observed 1.2M new users and 16.8M new links

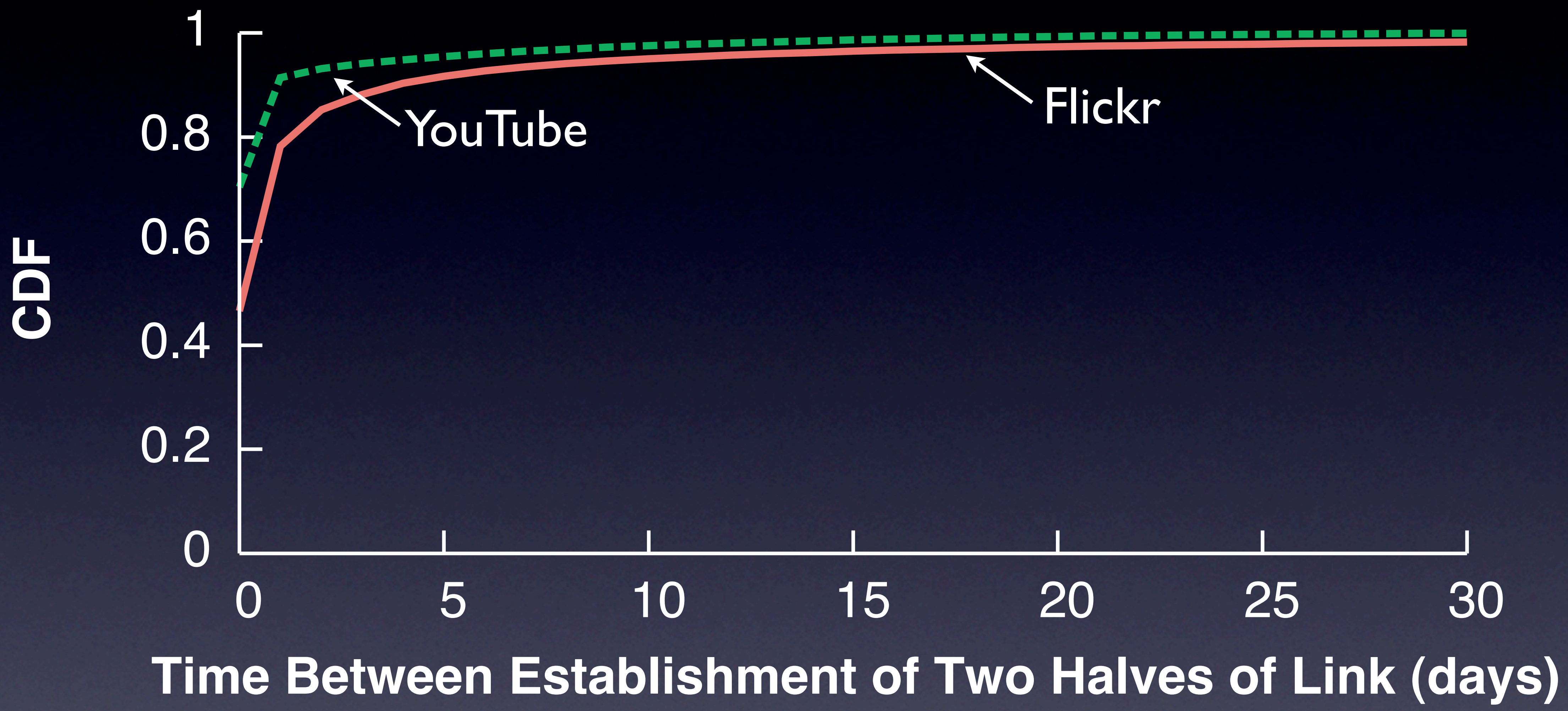
- Question: What processes are driving network growth?

# Growth data characteristics

| | Observation Period | Node Growth Rate | Link Growth Rate |
|---|---|---|---|
| Flickr | 104 days | 242% | 455% |
| YouTube | 36 days | 145% | 215% |
| Wikipedia | 825 days | 54% | 120% |
| Internet | 1,281 days | 31% | 43% |

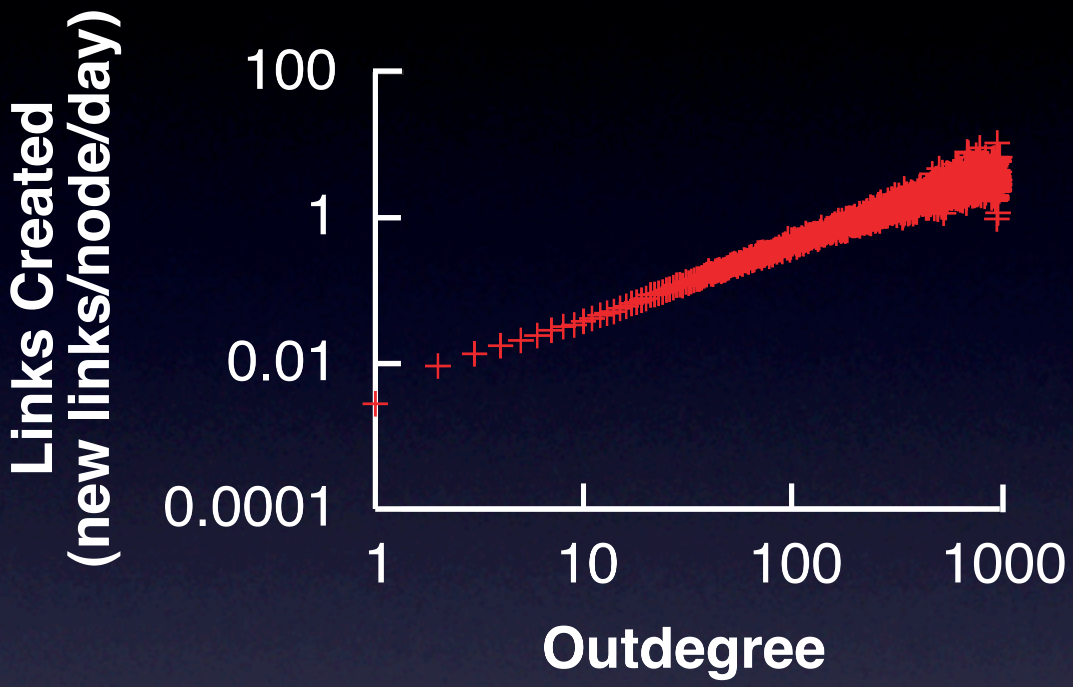
- Crawled social networks repeatedly for months
  - Observed 1.2M new users and 16.8M new links
- Question: What processes are driving network growth?

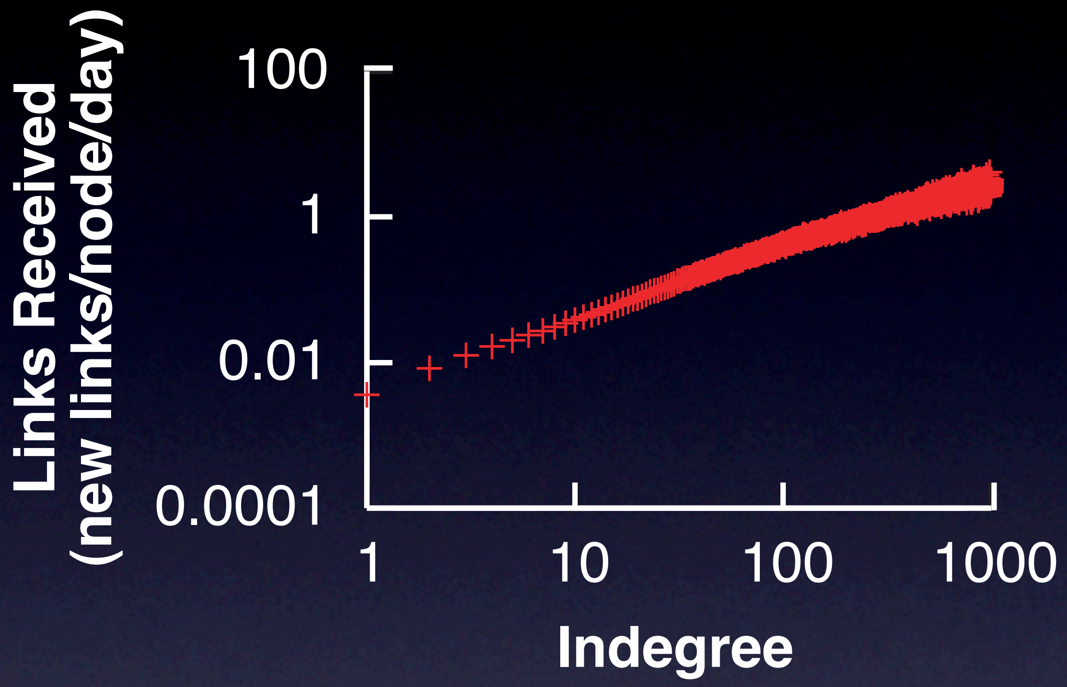
# Can *reciprocity* explain symmetry?



- Yes, over 80% of symmetric links created within 48 hours
  - Sites often inform users of new incoming links

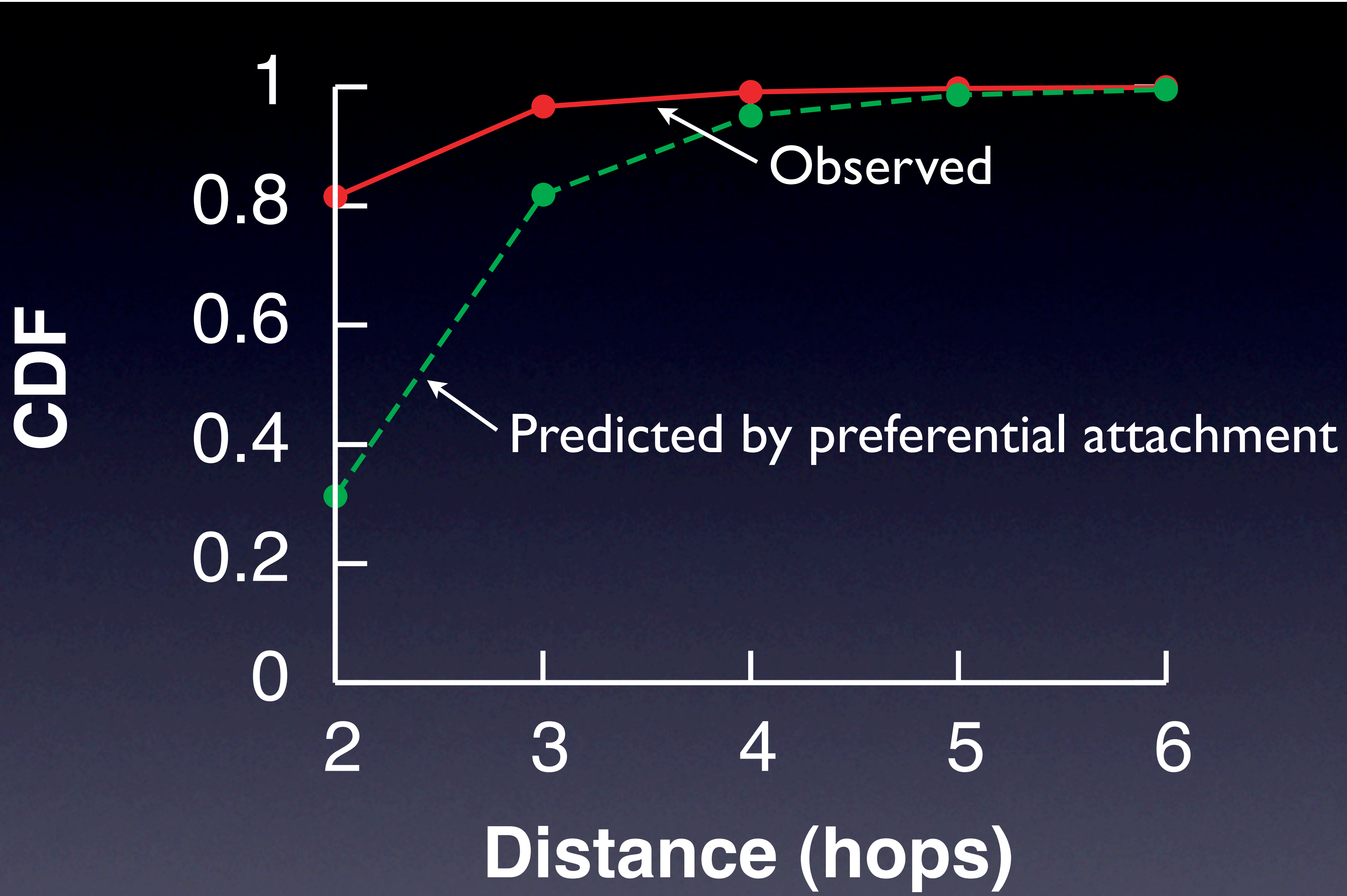
# Who creates and receives new links?





- Links created in proportion to outdegree (preferential creation)
- Links received in proportion to indegree (preferential reception)
- Is this preferential attachment?

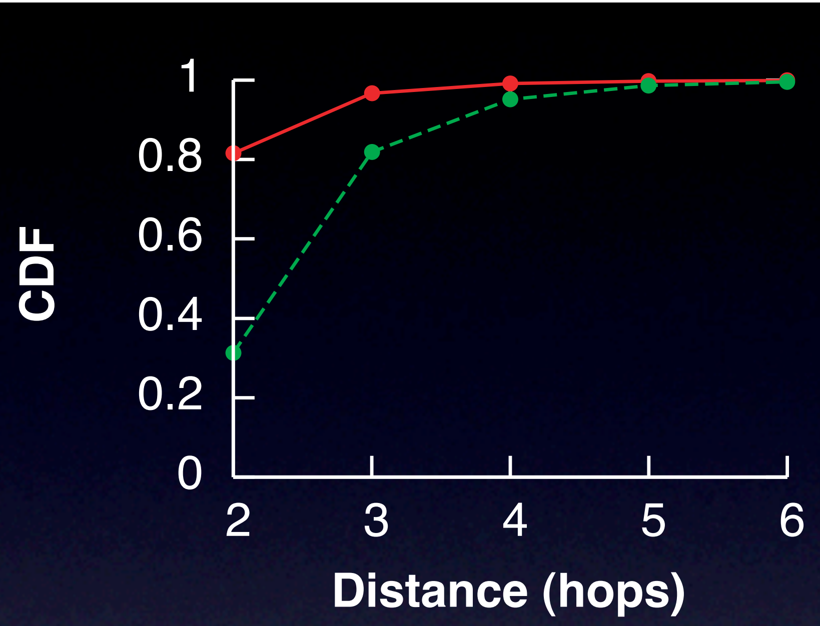
# Does proximity matter?

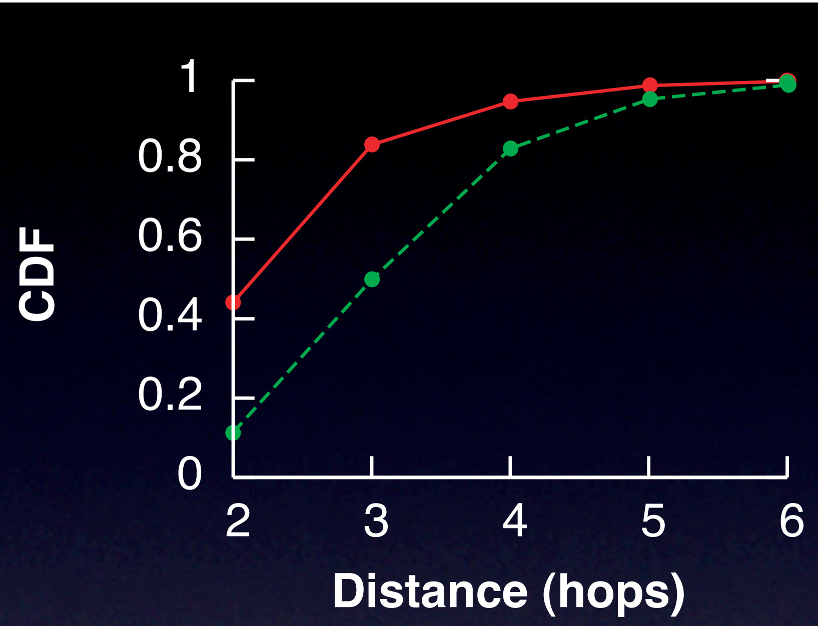


- New friends much closer than preferential attachment predicts
  - Suggests links created by local rules

# Does proximity matter?

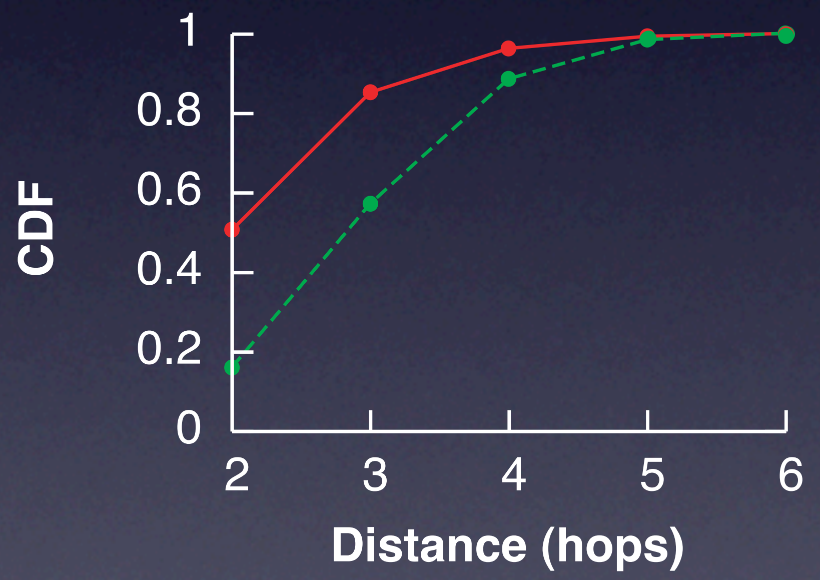
Flickr

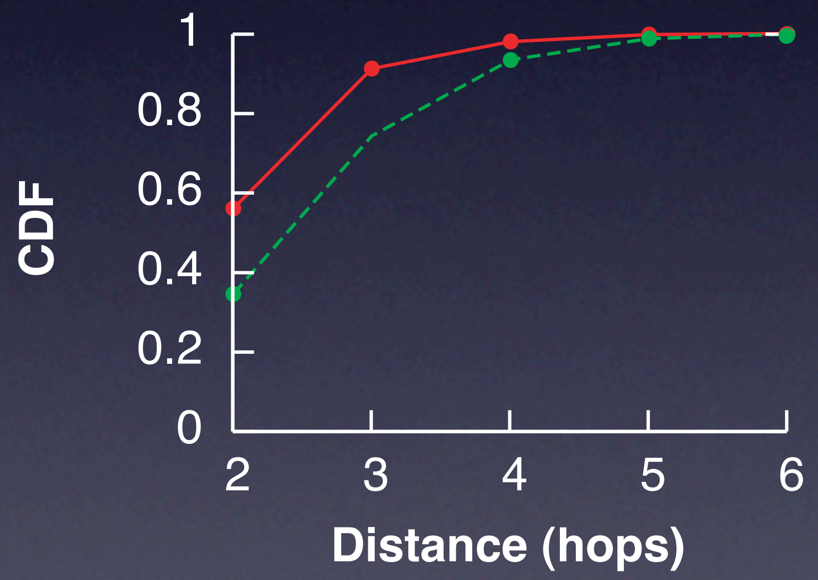




YouTube

Wikipedia





Internet

- New friends much closer than preferential attachment predicts
    - Suggests links created by local rules

# Implications of network growth

- Observed growth of large, complex information networks
  - 2.9M new users and 24M new links

- Found multiple growth processes at work
  - Reciprocity leads to high symmetry
  - Proximity bias leads to high clustering

- Modeling complex network growth
  - Based on local rules
  - Can validate or invalidate models with detailed data
  - Allow verification of systems at arbitrary size

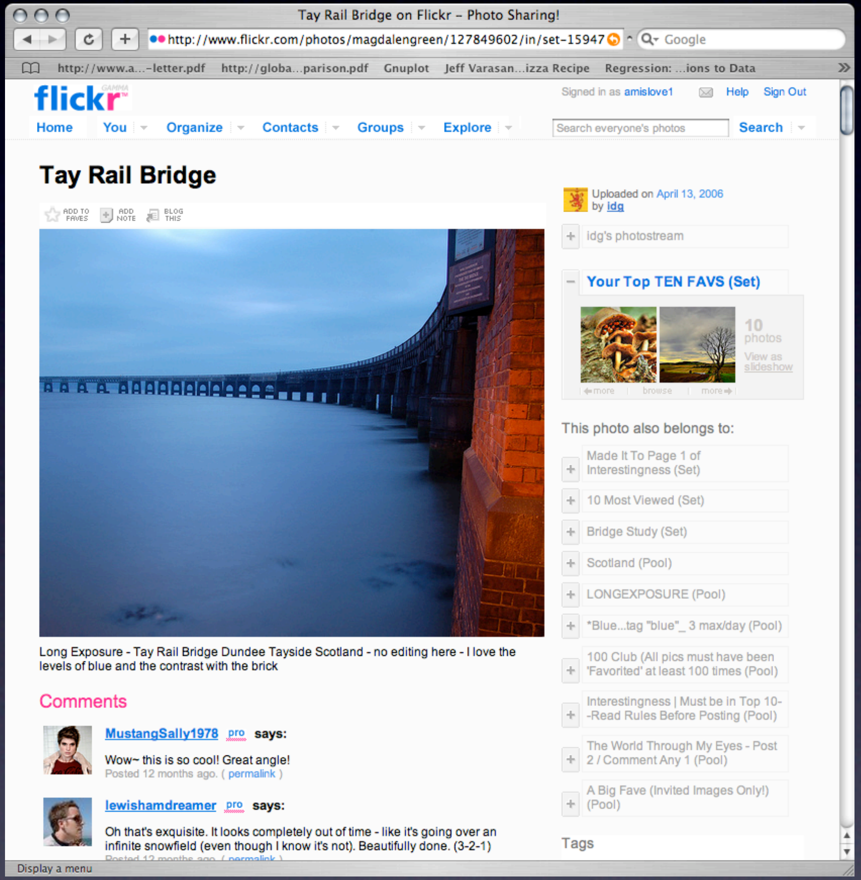
# Part III:

# Information flow

(ongoing work)

# Information flow

- Examining information flow in social networks
  - Lead to better information flow prediction and search algorithms
- Observe content propagating through network
  - Sample of 500,000 Flickr photos
- Examine different popularity metrics
  - Obtained history of *comments*, *favorites*
  - Recorded *views* daily



Tay Rail Bridge on Flickr – Photo Sharing!

http://www.flickr.com/photos/magdalengreen/127849602/in/set-15947

Tay Rail Bridge

Uploaded on April 13, 2006 by idg

This photo also belongs to:

- Made It To Page 1 of Interestingness (Set)
- 10 Most Viewed (Set)
- Bridge Study (Set)
- Scotland (Pool)
- LONGEXPOSURE (Pool)
- *Blue...tag "blue"_ 3 max/day (Pool)
- 100 Club (All pics must have been 'Favorited' at least 100 times (Pool)
- Interestingness | Must be in Top 10--Read Rules Before Posting (Pool)
- The World Through My Eyes - Post 2 / Comment Any 1 (Pool)
- A Big Fave (Invited Images Only!) (Pool)

Tags

Long Exposure - Tay Rail Bridge Dundee Tayside Scotland - no editing here - I love the levels of blue and the contrast with the brick

Comments

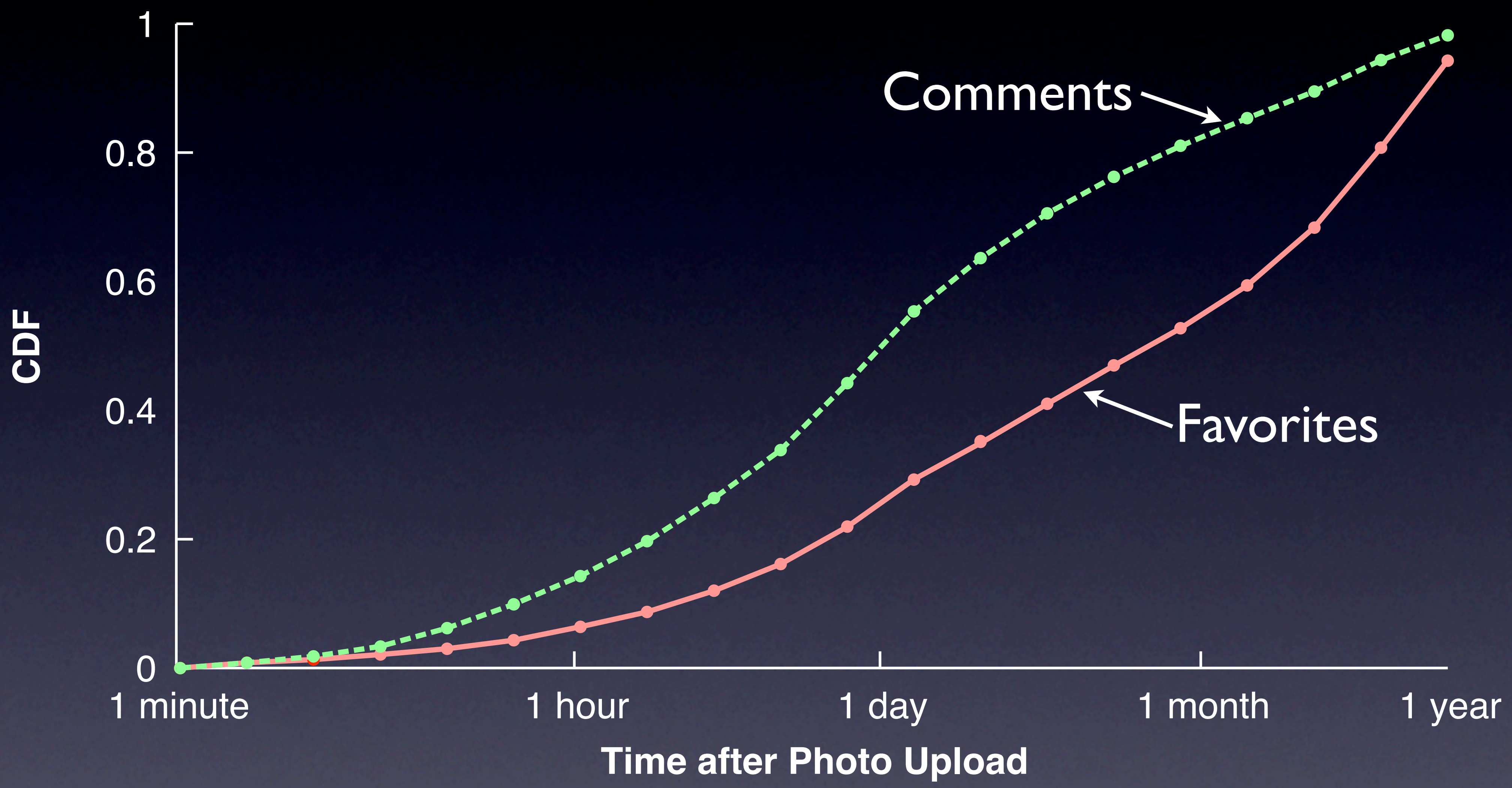
MustangSally1978 pro says:

Wow~ this is so cool! Great angle!
Posted 12 months ago. ( permalink )

lewishamdreamer pro says:

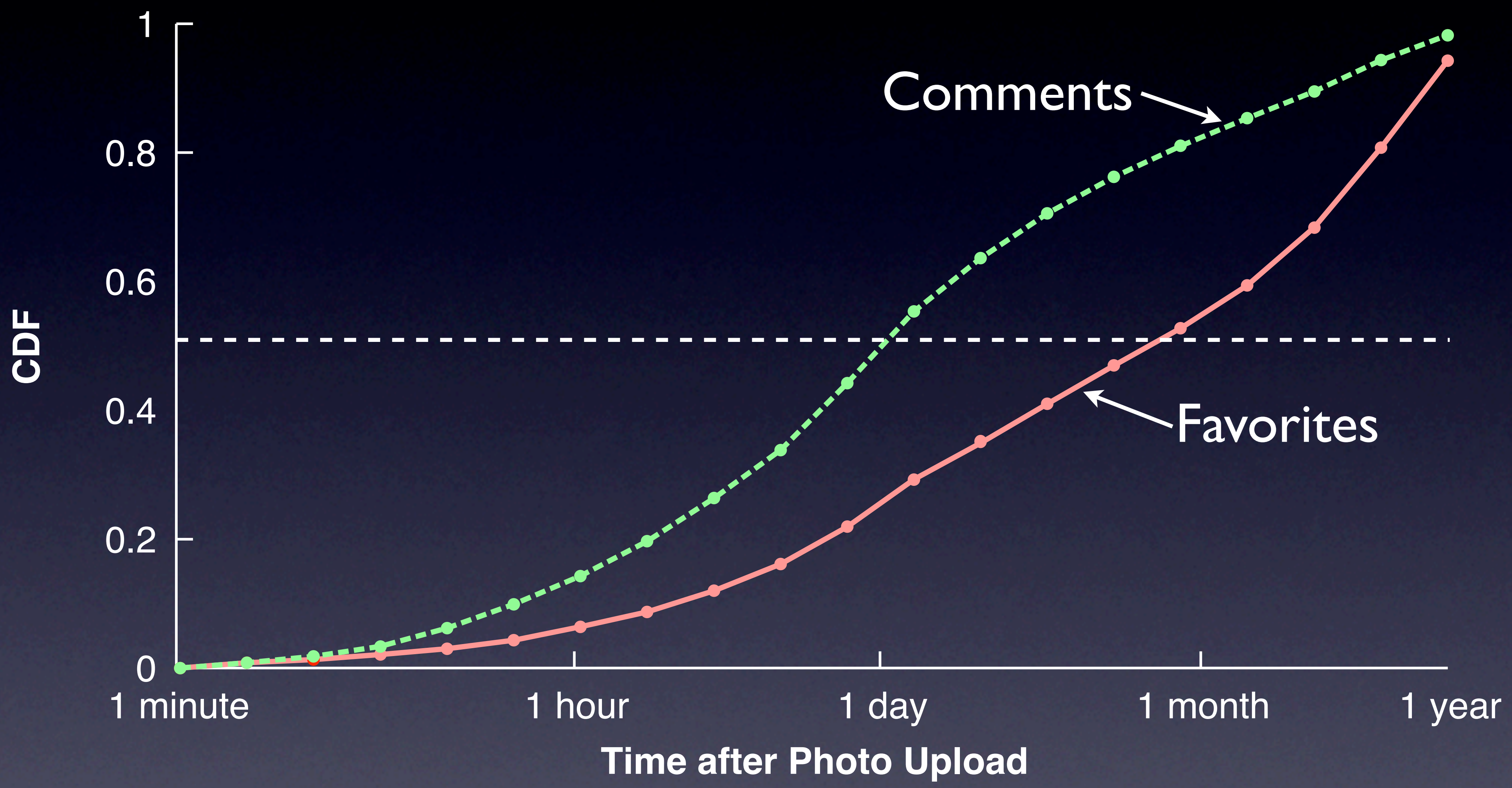
Oh that's exquisite. It looks completely out of time - like it's going over an infinite snowfield (even though I know it's not). Beautifully done. (3-2-1)
Posted 12 months ago. ( permalink )

# How quickly does content spread?



- 50 % of comments placed within 24 hours

# How quickly does content spread?



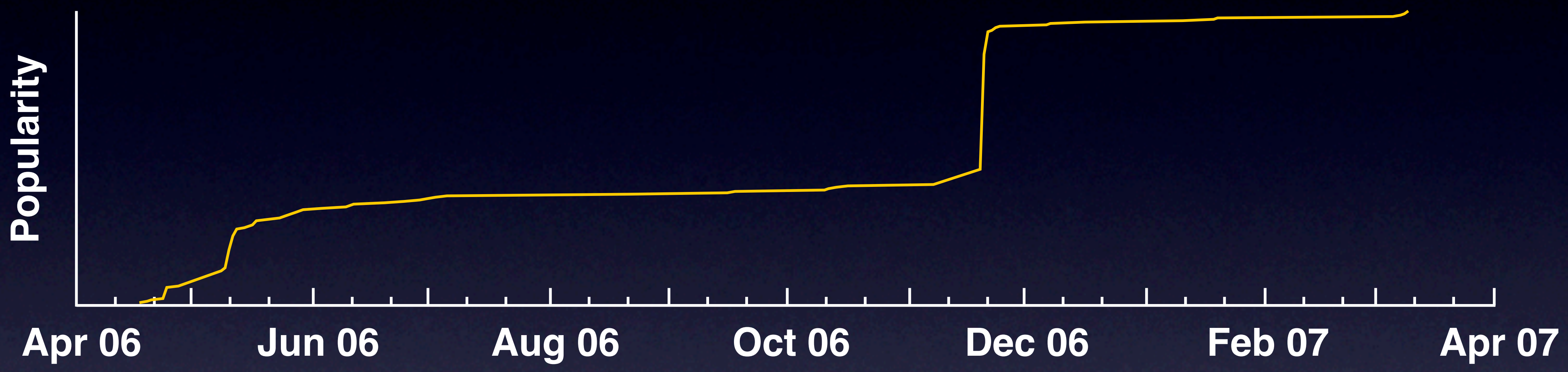
- 50 % of comments placed within 24 hours

# Rapid information propagation

- Information often propagates along links
  - Can track propagation

- What enables such rapid information flow?

# Rapid information propagation



- Information often propagates along links
    - Can track propagation
- What enables such rapid information flow?

# Summary

- Analyzed network structure and dynamics
  - Multiple networks have similar, unique characteristics
  - Consistent growth characteristics
- Many future directions
  - Examining information flow
  - Building better systems
- Open to expertise from other areas

**`http://socialnetworks.mpi-sws.org`**