Social network analysis with Hadoop

Jake Hofman

Yahoo! Research

October 2, 2009

Social networks

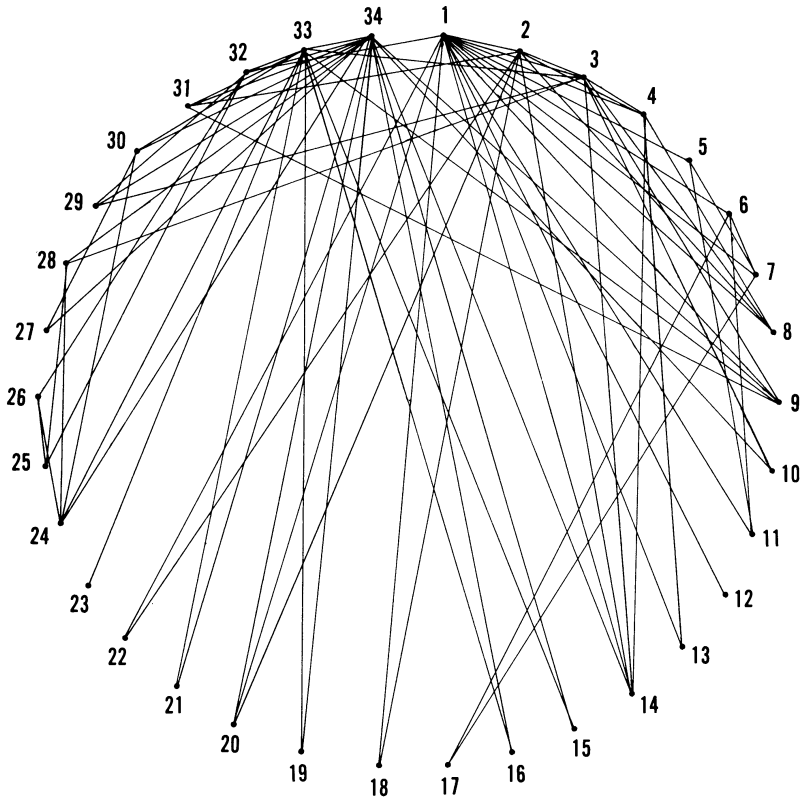
- Rapid increase in amount and variety of social network data

- Valuable information for products (recommendations, advertising, etc.) and research (structure/dynamics, diffusion, etc.)

Social networks

Goal: to enable analysis of large-scale social network data with readily available software/hardware

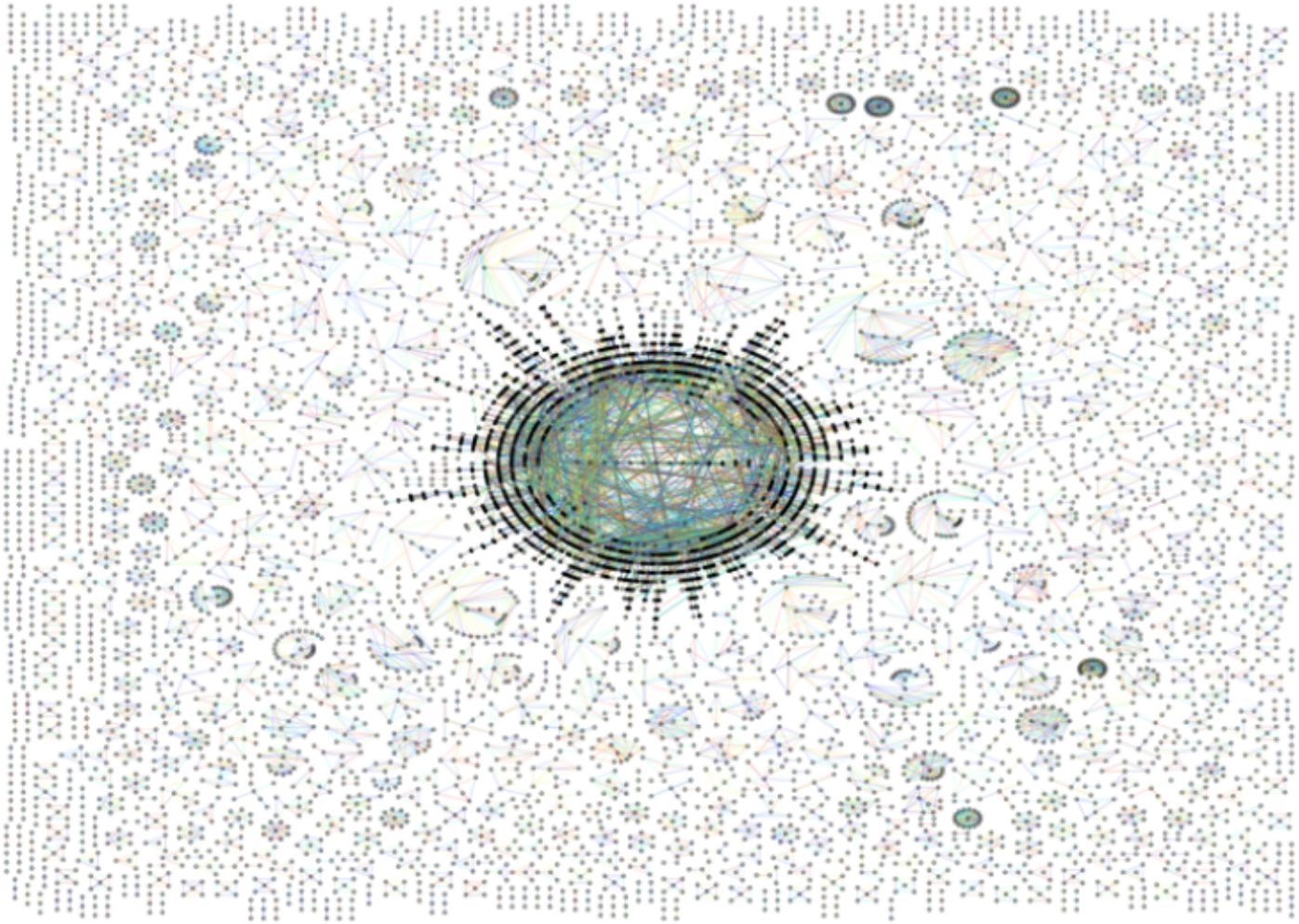
1970s $\sim 10^1$ nodes



The karate club was observed for a period of three years, from 1970 to 1972. In addition to direct observation, the history of the club prior to the period of the study was reconstructed through informants and club records in the university archives. During the period of observation, the

- Few direct observations; highly detailed info on nodes and edges
- E.g. karate club (Zachary, 1977)

1990s $\sim 10^4$ nodes



- Larger, indirect samples; relatively few details on nodes and edges
- E.g. APS co-authorship network (`http://bit.ly/aps08jmh`)

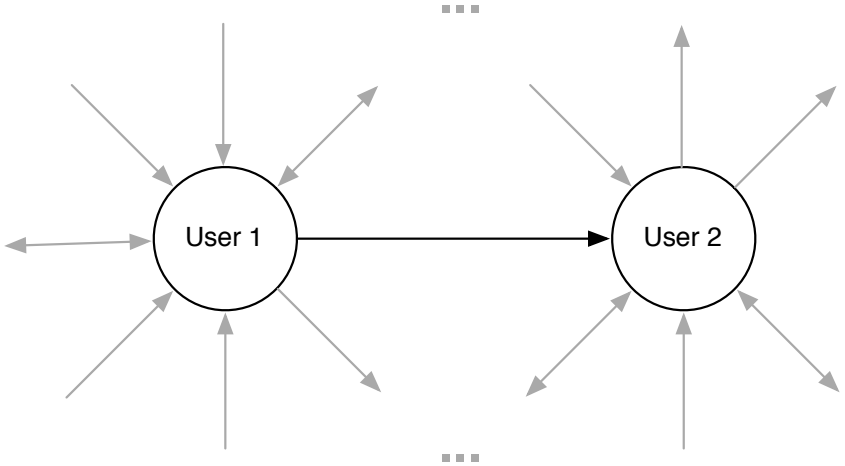
Present $\sim 10^8$ nodes +


```
<?xml version="1.0" encoding="UTF-8"?>
<user>
  <id>18805477</id>
  <name>jake hofman</name>
  <screen_name>jakehofman</screen_name>
  <location>new york, ny</location>
  <description>research scientist interested in machine learning for large data sets, including network, text and image data</description>
  <profile_image_url>http://a3.twimg.com/profile_images/71044209/jmh_dot_normal.jpg</profile_image_url>
  <url>http://www.jakehofman.com</url>
  <protected>false</protected>
  <followers_count>292</followers_count>
  <friends_count>208</friends_count>
  <created_at>Fri Jan 09 16:20:08 +0000 2009</created_at>
  <favourites_count>823</favourites_count>
  <statuses_count>628</statuses_count>
  ...
  <status>
    <created_at>Thu Oct 01 23:46:06 +0000 2009</created_at>
    <id>4538350570</id>
    <text>RT @atveit: RT @atbrox Mapreduce &amp; Hadoop Algorithms in Academic Papers - http://bit.ly/2rPgG #hadoopworld</text>
    <source>&lt;a href=&quot;http://www.atebits.com/&quot; rel=&quot;nofollow&quot;&gt;Tweetie&lt;/a&gt;</source>
    <truncated>false</truncated>
    <in_reply_to_status_id></in_reply_to_status_id>
    <in_reply_to_user_id></in_reply_to_user_id>
    <favorited>false</favorited>
    <in_reply_to_screen_name></in_reply_to_screen_name>
    <geo/>
  </status>
</user>
```


- Very large, dynamic samples; many details in node and edge metadata
- E.g. Mail, Messenger, Facebook, Twitter, etc.

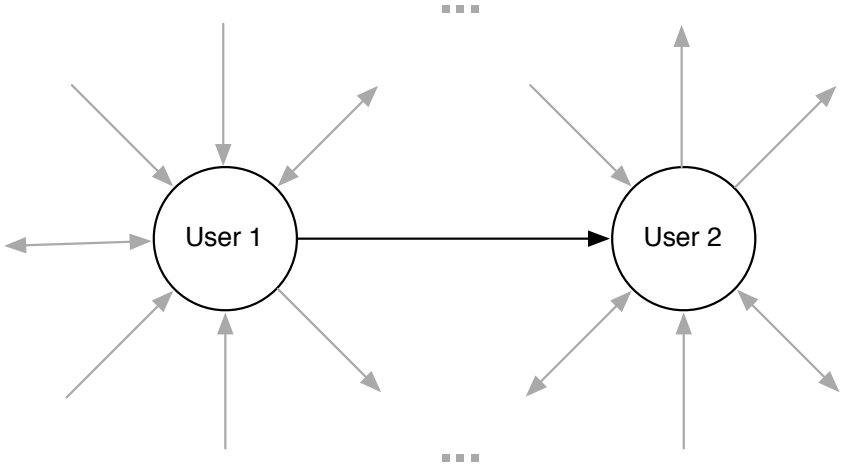
Scale

- Example numbers:
  - $\sim 10^7$ nodes
  - $\sim 10^2$ edges/node (degree)
  - no node/edge data
  - static
  - ~8GB



Scale

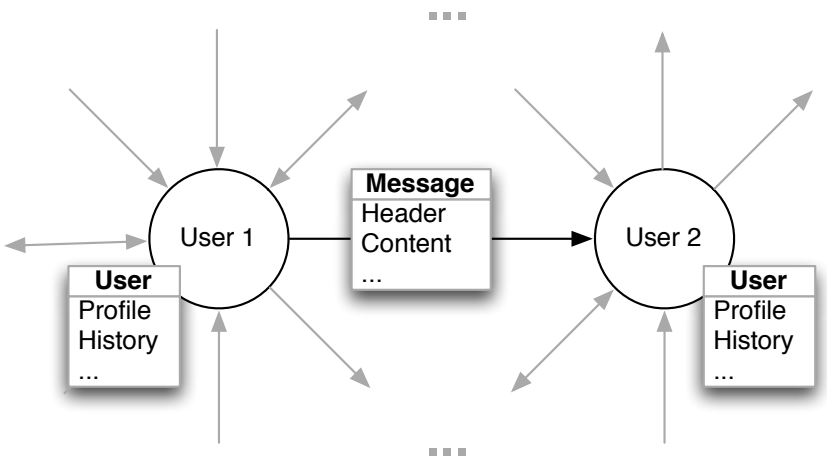
- Example numbers:
  - $\sim 10^7$ nodes
  - $\sim 10^2$ edges/node (degree)
  - no node/edge data
  - static
  - ~8GB



Simple, static networks push memory limit for commodity machines

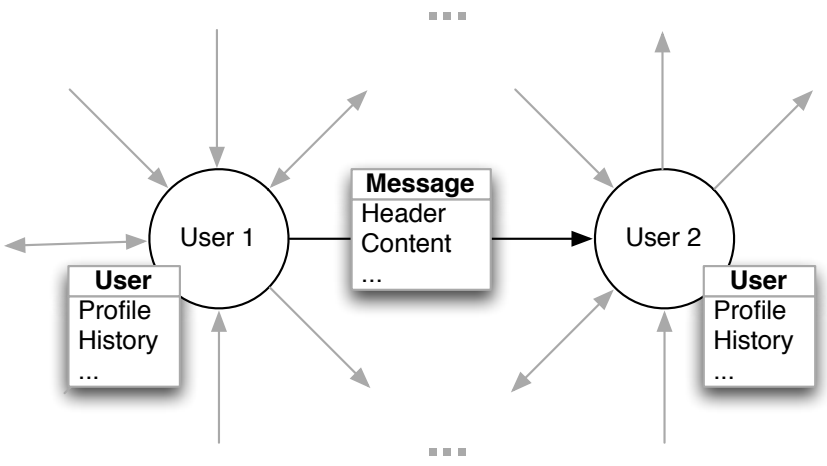
Scale

- Example numbers:
  - $\sim 10^7$ nodes
  - $\sim 10^2$ edges/node (degree)
  - node/edge metadata
  - dynamic
  - $\sim$100GB/day



Scale

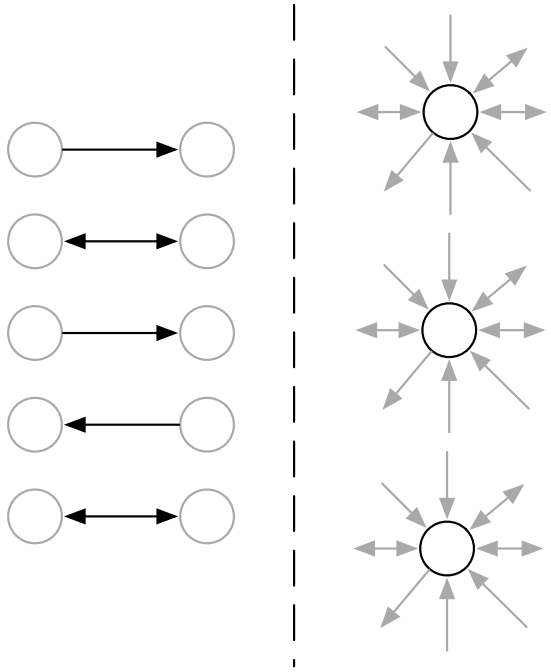
- Example numbers:
  - $\sim 10^7$ nodes
  - $\sim 10^2$ edges/node (degree)
  - node/edge metadata
  - dynamic
  - $\sim$100GB/day



Dynamic, data-rich social networks exceed memory limits; require considerable storage

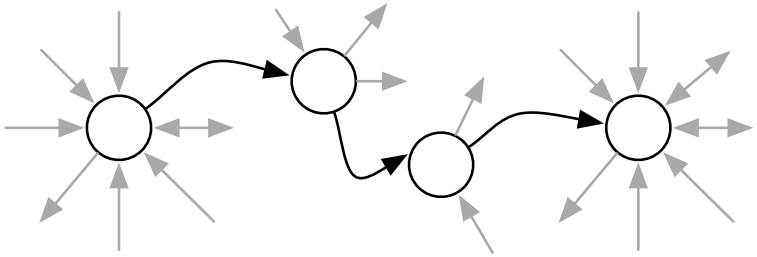
Distributed network analysis

MapReduce convenient for parallelizing individual node/edge-level calculations



Distributed network analysis

Higher-order calculations more difficult when network exceeds memory constraints, but can be adapted to MapReduce framework



Package details

- Network creation/manipulation
  - Logs $\rightarrow$ edges
  - Edge list $\leftrightarrow$ adjacency list
  - Directed $\leftrightarrow$ undirected
  - Edge thresholds
- First-order descriptive statistics
  - Number of nodes
  - Number of edges
  - Node degrees
- Higher-order node-level descriptive statistics
  - Clustering coefficient
  - Implicit degree
  - ...
- Global calculations
  - Pairwise connectivity
  - Connected components
  - Minimum spanning tree
  - Breadth-first search
  - Pagerank
  - Community detection

Package details

- Network creation/manipulation
  - Logs $\rightarrow$ edges
  - Edge list $\leftrightarrow$ adjacency list
  - Directed $\leftrightarrow$ undirected
  - Edge thresholds
- First-order descriptive statistics
  - Number of nodes
  - Number of edges
  - Node degrees
- Higher-order node-level descriptive statistics
  - Clustering coefficient
  - Implicit degree
  - ...
- Global calculations
  - Pairwise connectivity
  - Connected components
  - Minimum spanning tree
  - Breadth-first search
  - Pagerank
  - Community detection

Currently implemented in Streaming with Python

Algorithms exist/developed for additional features

Application: Twitter


```
<?xml version="1.0" encoding="UTF-8"?>
<user>
  <id>19041500</id>
  <name>Apache Hadoop</name>
  <screen_name>hadoop</screen_name>
  <location></location>
  <description>Open source software for reliable, distributed, scalable computing.</description>
  <profile_image_url>http://a3.twimg.com/profile_images/72618237/ScreenShot002_normal.png</profile_image_url>
  <url>http://hadoop.apache.org</url>
  <protected>false</protected>
  <followers_count>703</followers_count>
  ...
  <friends_count>6</friends_count>
  <created_at>Thu Jan 15 22:05:30 +0000 2009</created_at>
  <favourites_count>0</favourites_count>
  <statuses_count>17</statuses_count>
  <following></following>
  <status>
    <created_at>Mon Jun 01 18:09:11 +0000 2009</created_at>
    <id>1993932474</id>
    <text>RT @cloudera: Database import for Hadoop. Details on Sqoop from Aaron Kimball: http://bit.ly/GQQTw</text>
    <source>web</source>
    <truncated>false</truncated>
    <in_reply_to_status_id></in_reply_to_status_id>
    <in_reply_to_user_id></in_reply_to_user_id>
    <favorited>false</favorited>
    <in_reply_to_screen_name></in_reply_to_screen_name>
    <geo/>
  </status>
</user>
```

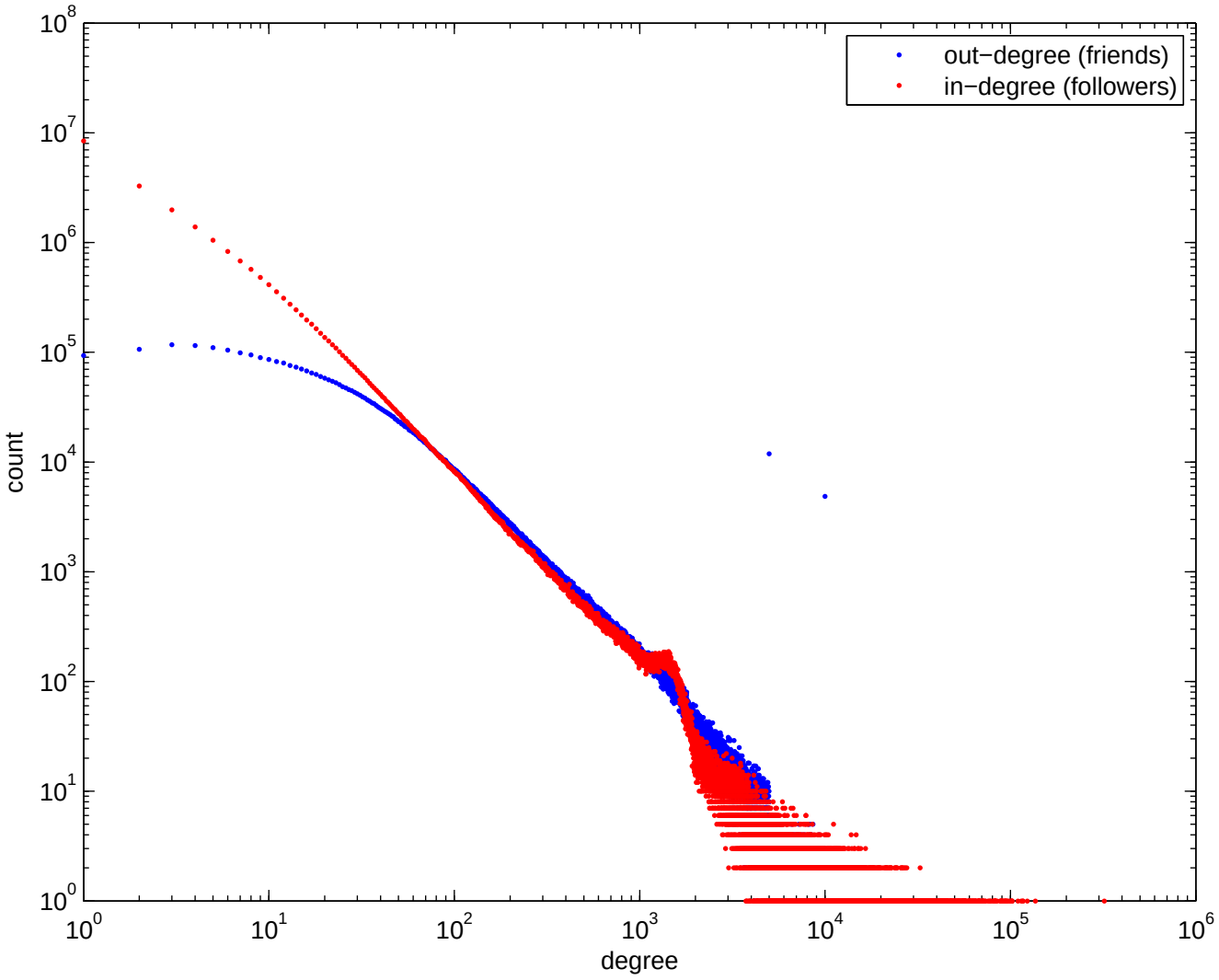

- Distributed crawl of Twitter social network + public messages (crawler by Eytan Bakshy, `http://bit.ly/eytanb`)

Application: Twitter


```
<?xml version="1.0" encoding="UTF-8"?>
<user>
  <id>19041500</id>
  <name>Apache Hadoop</name>
  <screen_name>hadoop</screen_name>
  <location></location>
  <description>Open source software for reliable, distributed, scalable computing.</description>
  <profile_image_url>http://a3.twimg.com/profile_images/72618237/ScreenShot002_normal.png</profile_image_url>
  <url>http://hadoop.apache.org</url>
  <protected>false</protected>
  <followers_count>703</followers_count>
  ...
  <friends_count>6</friends_count>
  <created_at>Thu Jan 15 22:05:30 +0000 2009</created_at>
  <favourites_count>0</favourites_count>
  <statuses_count>17</statuses_count>
  <following></following>
  <status>
    <created_at>Mon Jun 01 18:09:11 +0000 2009</created_at>
    <id>1993932474</id>
    <text>RT @cloudera: Database import for Hadoop. Details on Sqoop from Aaron Kimball: http://bit.ly/GQQTw</text>
    <source>web</source>
    <truncated>false</truncated>
    <in_reply_to_status_id></in_reply_to_status_id>
    <in_reply_to_user_id></in_reply_to_user_id>
    <favorited>false</favorited>
    <in_reply_to_screen_name></in_reply_to_screen_name>
    <geo/>
  </status>
</user>
```

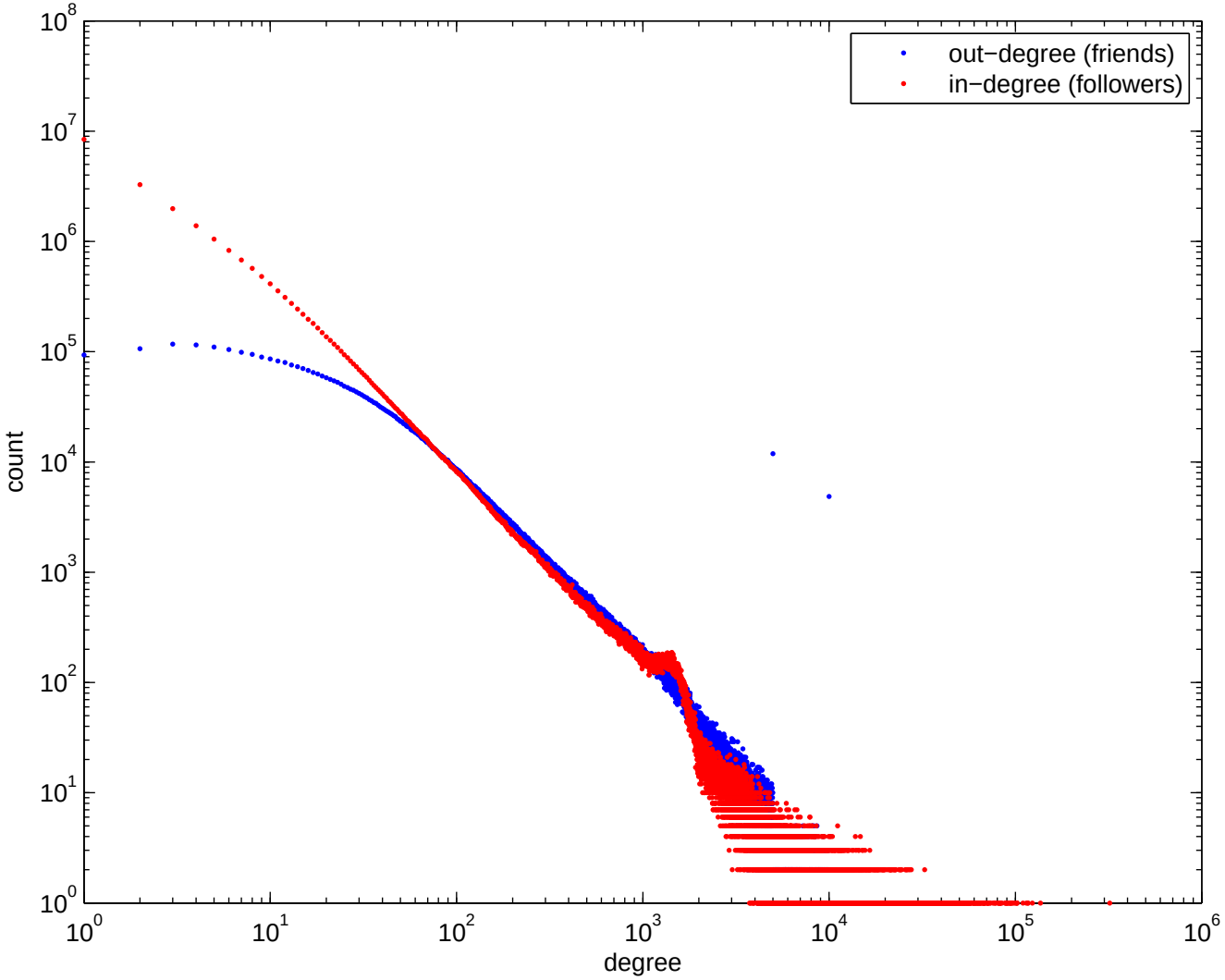

- Distributed crawl of Twitter social network + public messages (crawler by Eytan Bakshy, `http://bit.ly/eytanb`)
- $\sim$ 25 million nodes, $\sim$ 800 million edges

Twitter: Degree Distribution



- Aggregates users by number of friends/followers seen in crawl

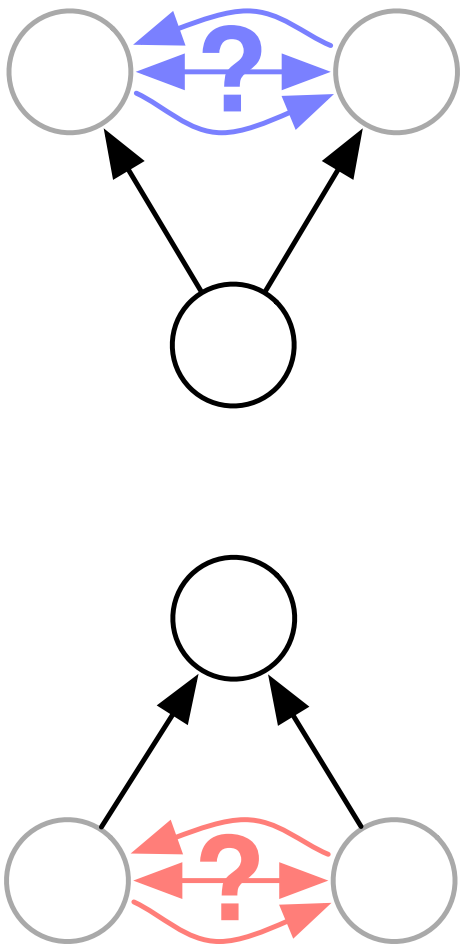
Twitter: Degree Distribution



Many people not followed by anyone; few followed by many

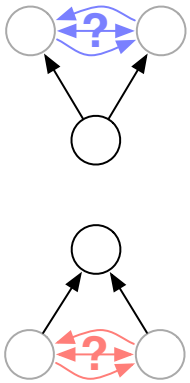
Most people follow at least a few others

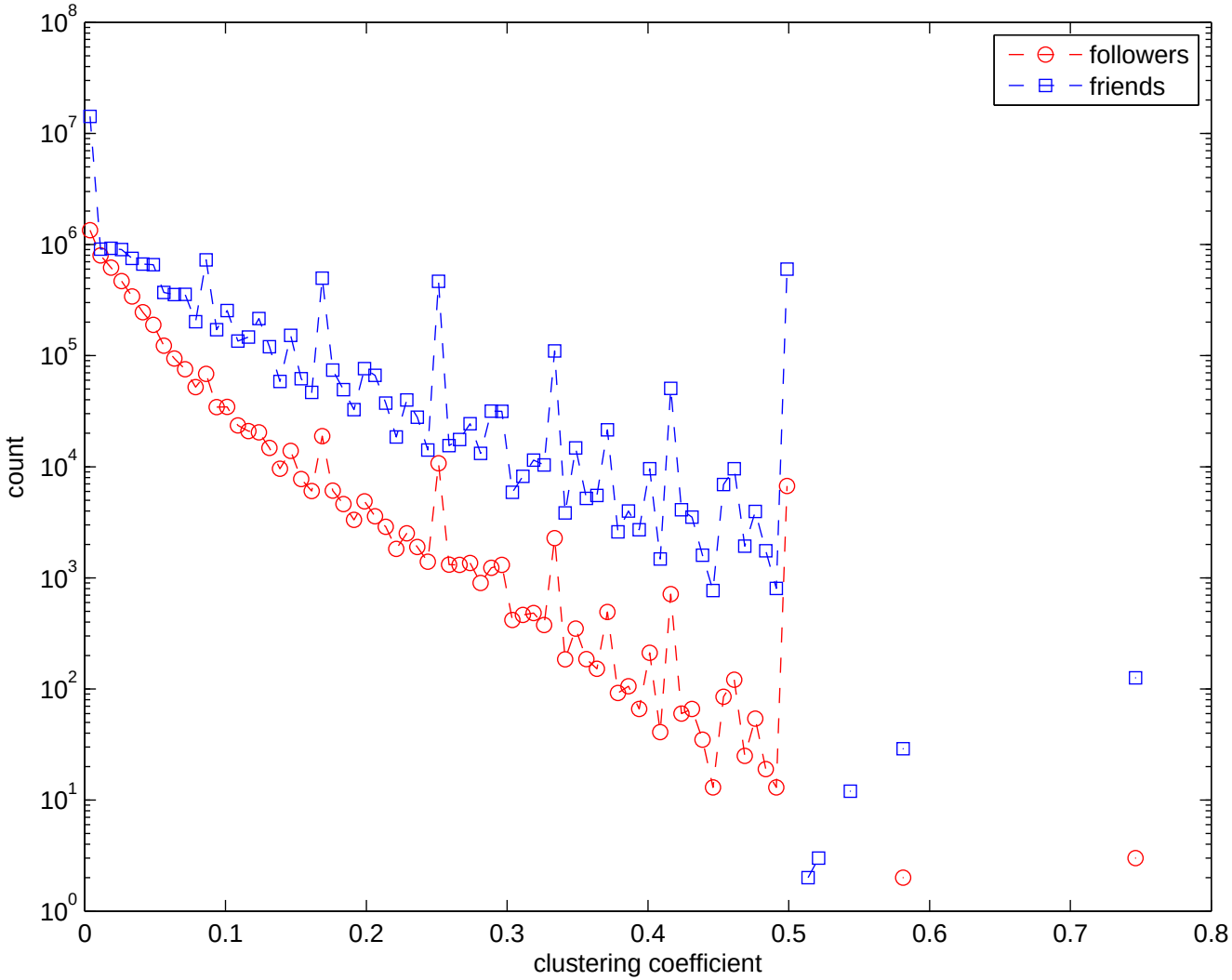
Twitter: Node-level clustering coefficient



- Fraction of edges amongst a node's friends/followers (Watts & Strogatz, 1998)

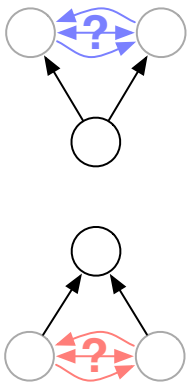
Twitter: Node-level clustering coefficient

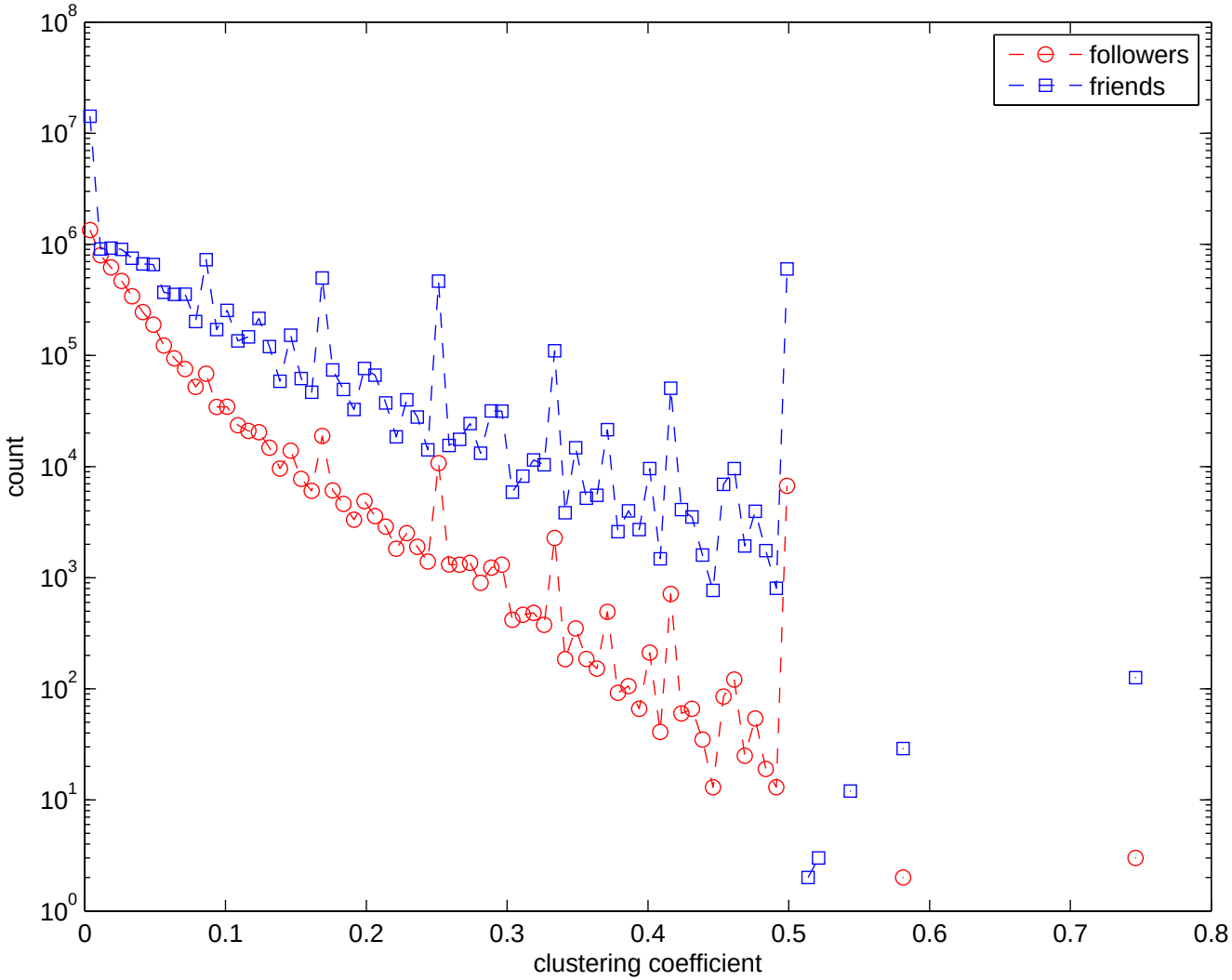




- Fraction of edges amongst a node's friends/followers (Watts & Strogatz, 1998)

Twitter: Node-level clustering coefficient





Suprisingly high density at 0.5 (many isolated triangles)

Future plans

- Open-source release
- *"A Model of Computation for MapReduce"*, Karloff, Suri, & Vassilvitskii, Symposium on Discrete Algorithms, 2010 (Accepted)
- Twitter analysis publication (In progress)

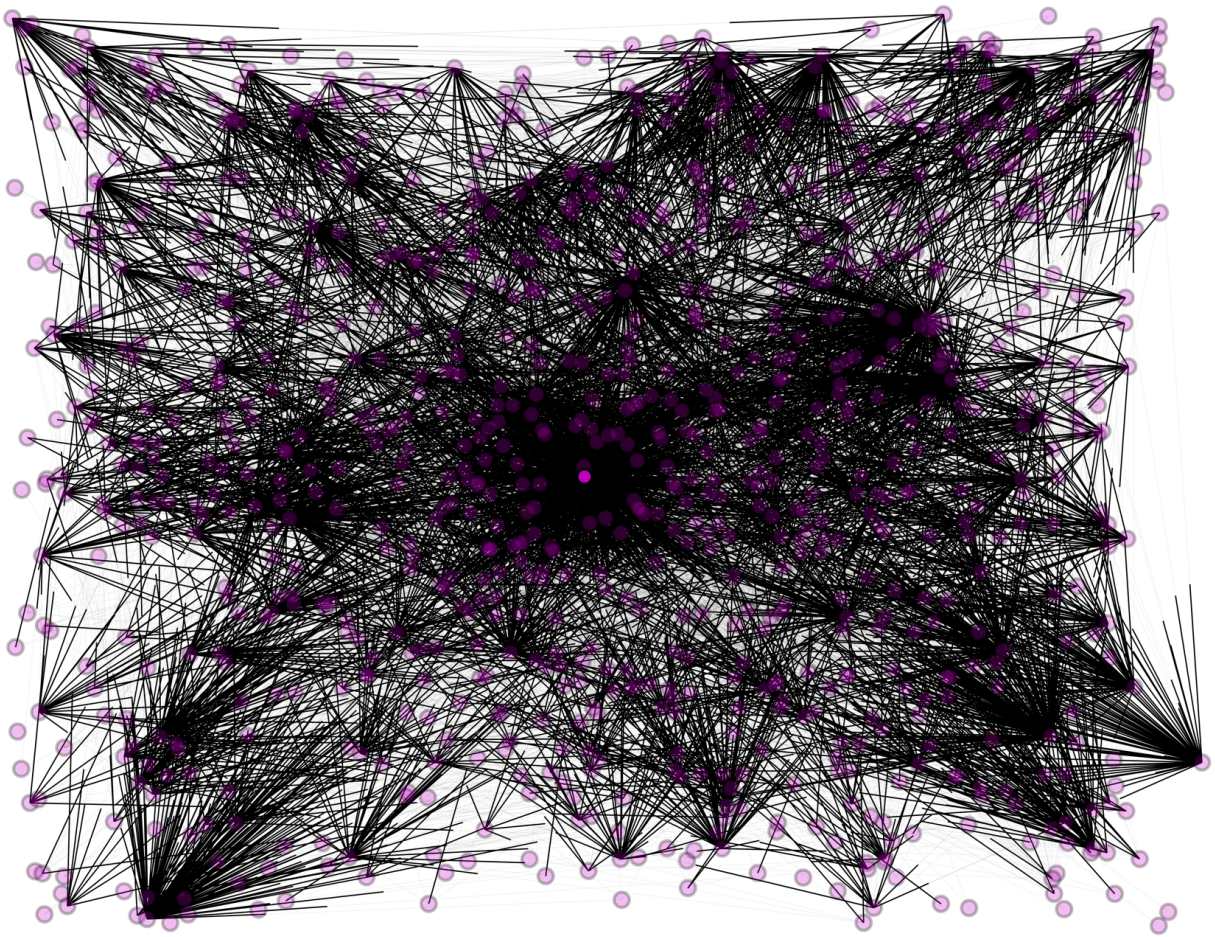
Goal: to enable analysis of large-scale social network data with readily available software/hardware

Collaborators

- Eytan Bakshy[m,y]
- Sharad Goel[y]
- Winter Mason[y]
- Sid Suri[y]
- Sergei Vassilvitskii[y]
- Duncan Watts[y]
- (You?)

[y] Yahoo! Research (`http://research.yahoo.com`)
[m] University of Michigan

Thanks.



Questions?[1]

[1] hofman@yahoo-inc.com, jakehofman.com