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.)
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
Present $\sim 10^8$ nodes +

- 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
  - $\sim 8$GB
Scale

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

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
  - $\sim 100\text{GB}/\text{day}$
Scale

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

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

Currently implemented in Streaming with Python

Algorithms exist/developed for additional features

Jake Hofman (Yahoo! Research)  
Social network analysis with Hadoop  
October 2, 2009
Package details

- Network creation/manipulation
  - Logs → edges
  - Edge list ↔ adjacency list
  - Directed ↔ 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

Application: Twitter

- ~ 25 million nodes, ~ 800 million edges
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)
• 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\textsuperscript{m,y}
- Sharad Goel\textsuperscript{y}
- Winter Mason\textsuperscript{y}
- Sid Suri\textsuperscript{y}
- Sergei Vassilvitskii\textsuperscript{y}
- Duncan Watts\textsuperscript{y}
- (You?)

\textsuperscript{y} Yahoo! Research (http://research.yahoo.com)
\textsuperscript{m} University of Michigan
Thanks.

Questions?¹

¹hofman@yahoo-inc.com, jakehofman.com