Data Sets for Large-Scale Social Network Analysis

Christopher DuBois
Ph.D. Student
Department of Statistics, UCI

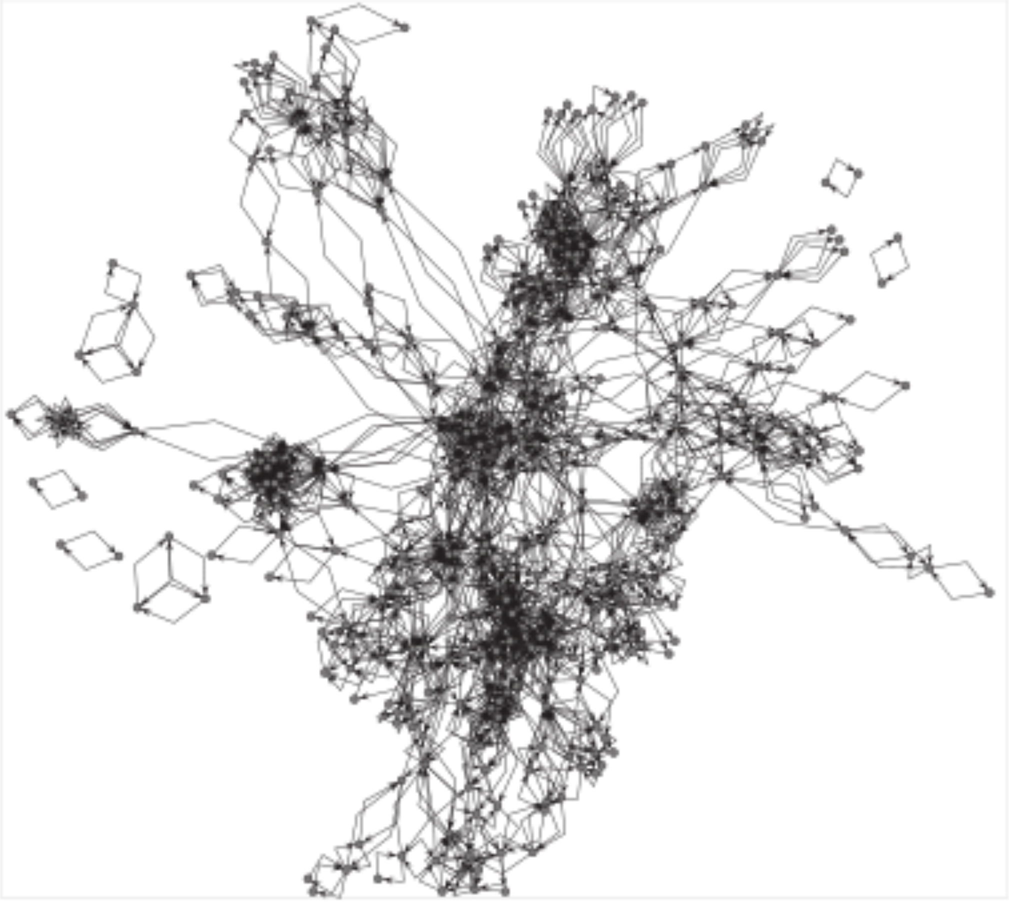
Overview

Data sets from online interaction

- Email, wikis, social networking sites, etc.
- Motivations for these data sets
- Problems applying current methods
- Opportunities

Enron Emails

- Collection of anonymized emails
- Over 1000 employees
- 2 years



O'Madadhain, Hutchins, Smyth. SIGKDD, 2004.

- Modeling challenges
- Baseline for comparison

Blogosphere Data

- 12 million blogs
- 40,000 posts per day

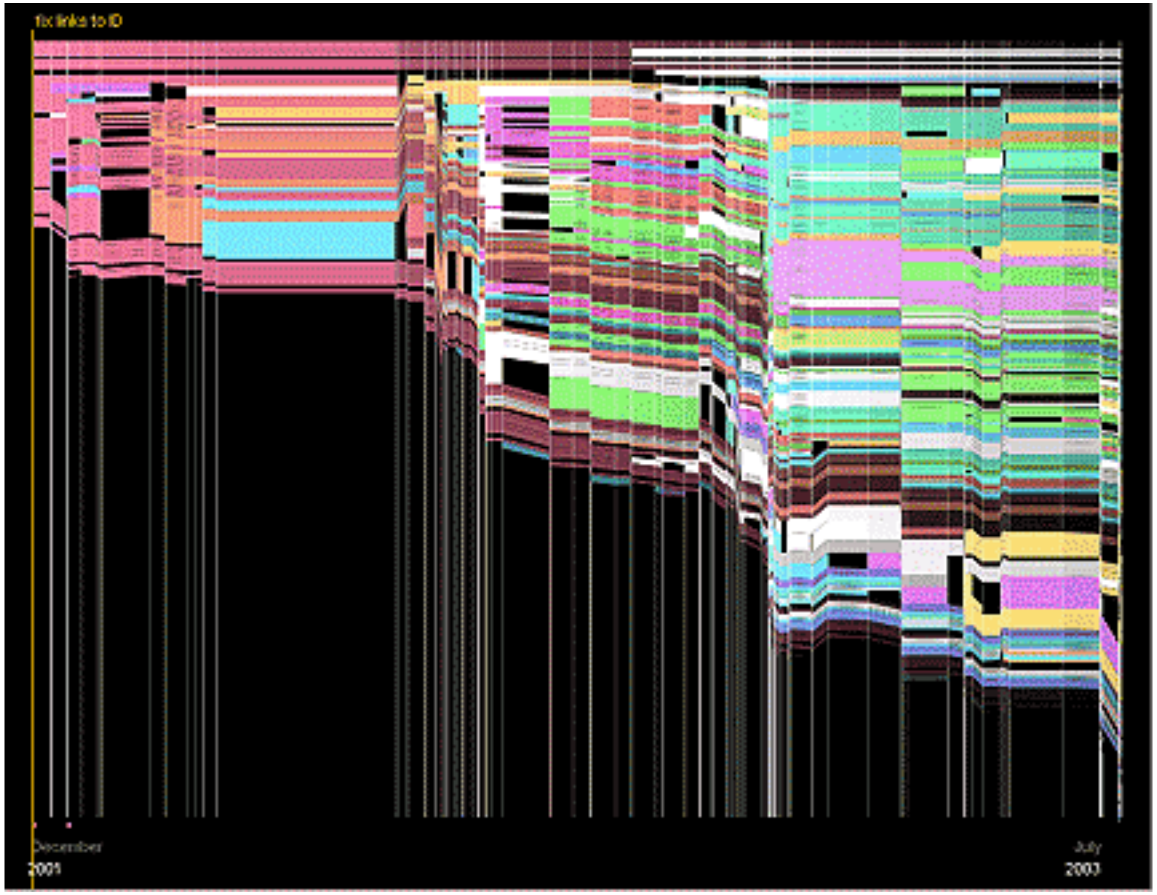
- Modeling challenges:
    - exact temporal info

Spinn3r.com

Wikipedia

- 10 million articles
- 5.8 million users
- (~500,000 highly active)
- Since 2001

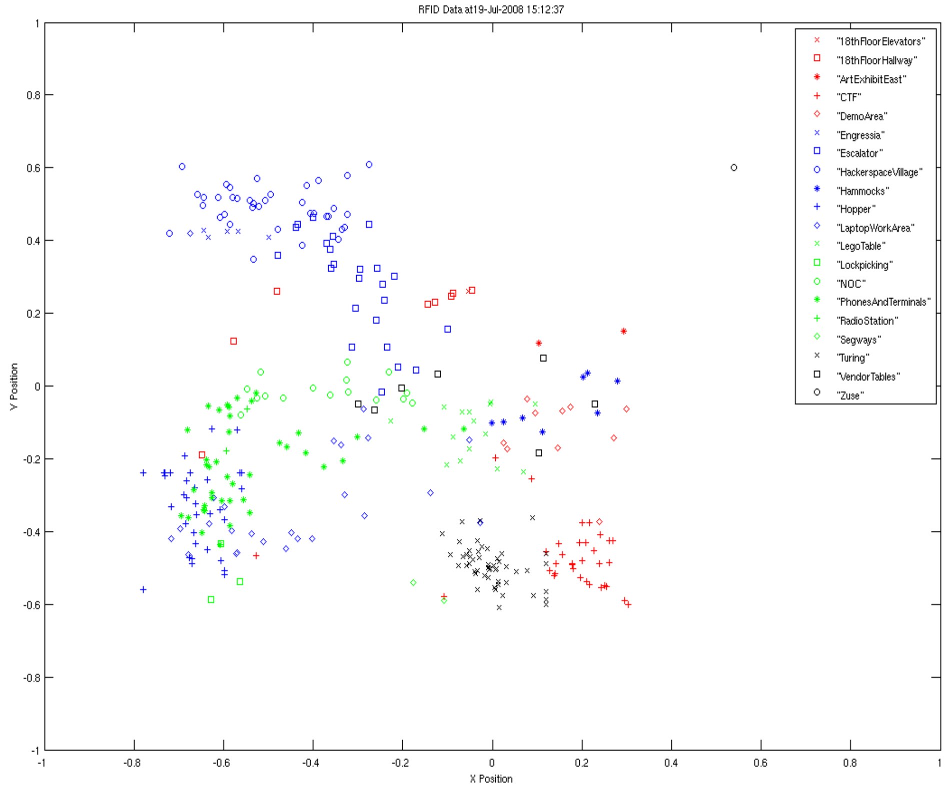
Not only relationships, but also text and time



Visualization for edits to the Wiki-article "Evolution" (IBM Research)

Conference Tracking

- 3 day conference
- 800 attendees
- position data every 30 seconds via RFID tags
- empirical face-to-face network

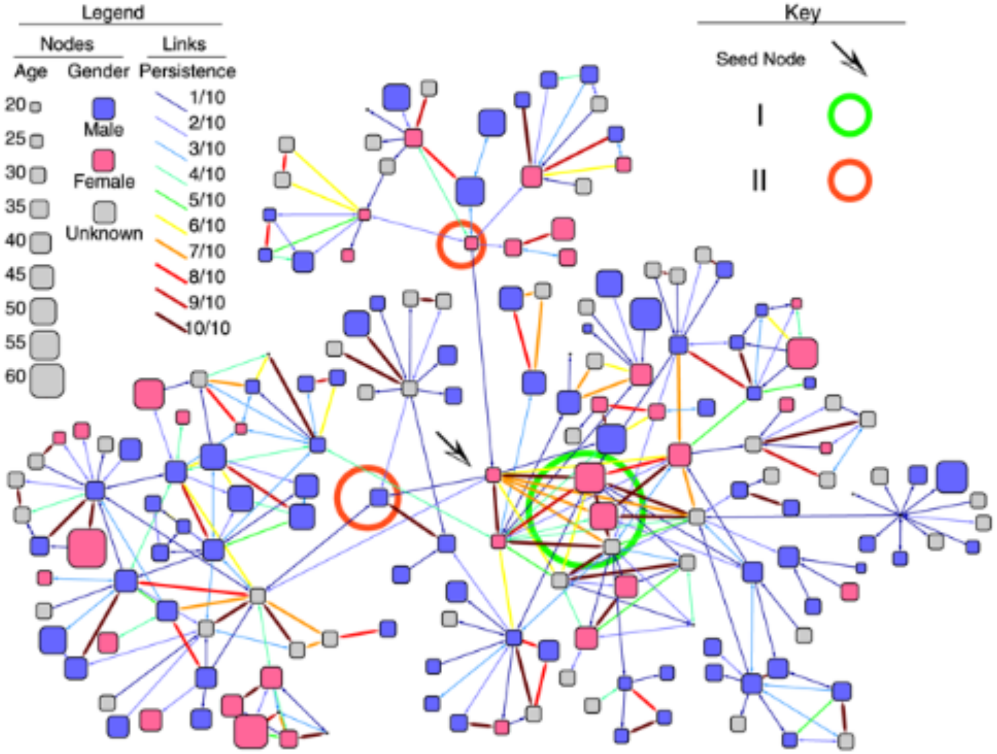


Additional attributes to consider

AMD Hope Team, 2008

Mobile phone calls

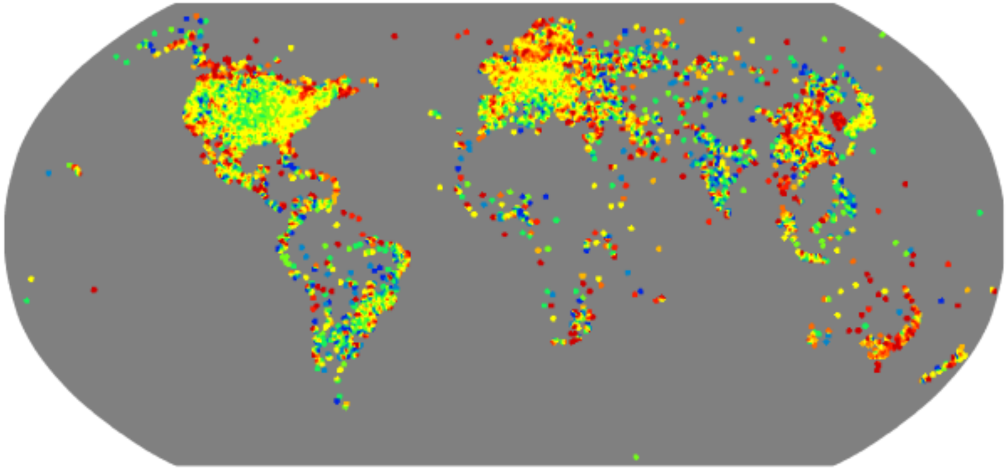
8,000,000 calls among more than 1,000,000 people.

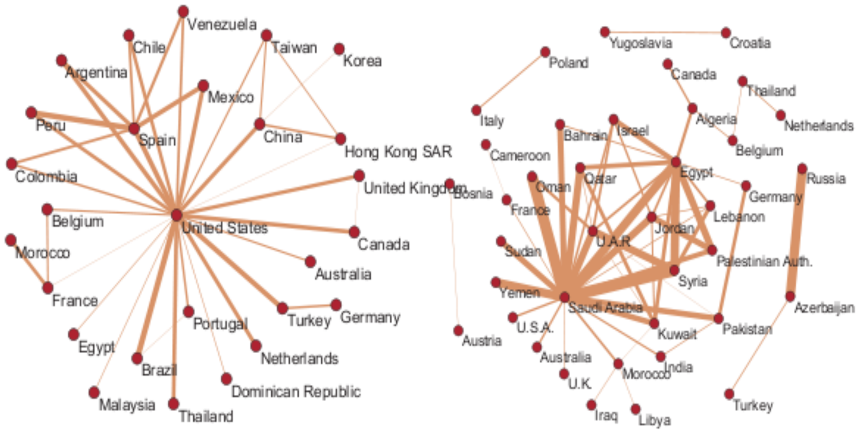


Hidalgo, Rodriguez-Sicker. The dynamics of a mobile phone network. Physica A 2007.

MSN Messenger

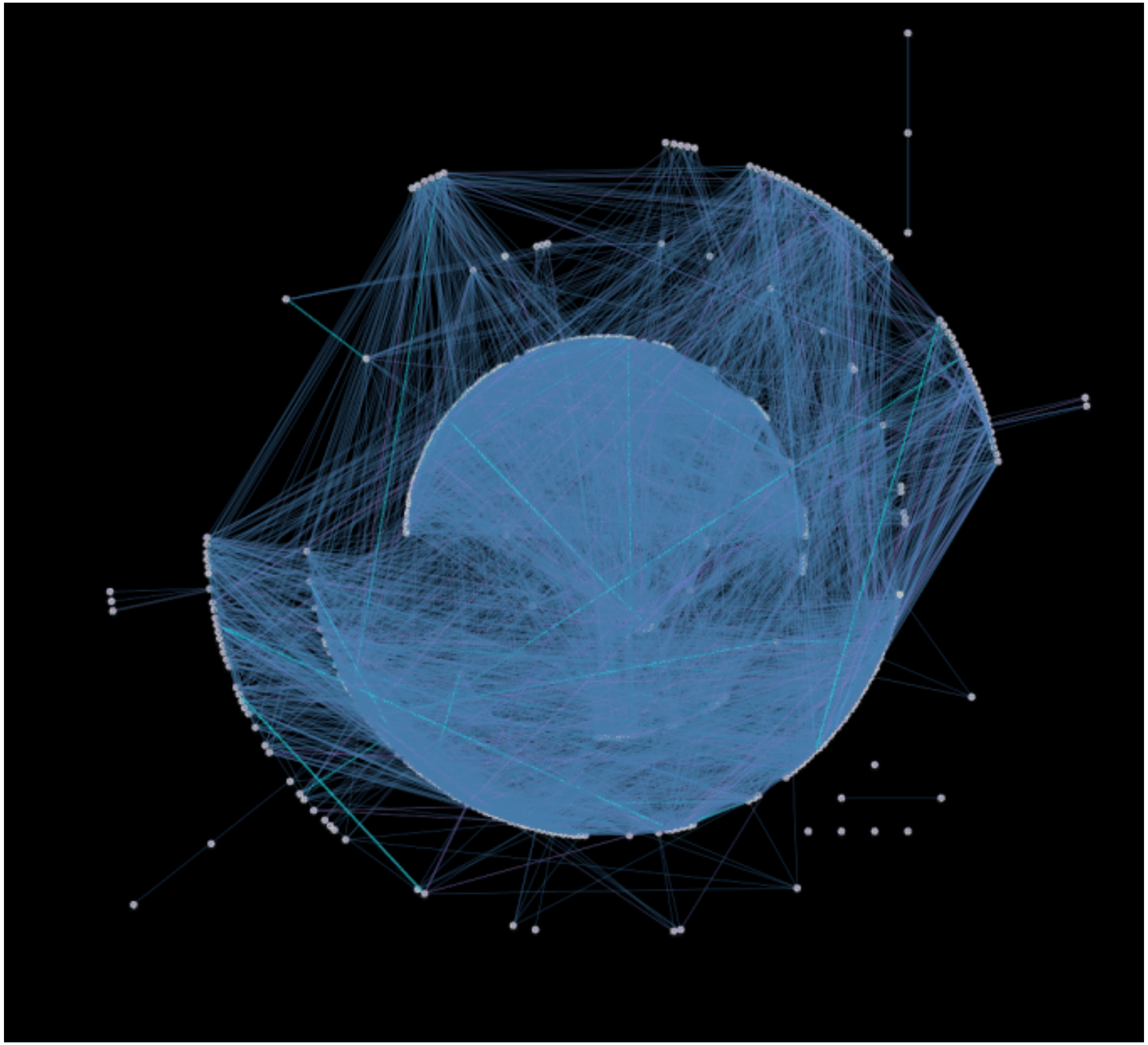
- Data from the month of June, 2006
- 180 million users
- 1.3 billion conversations





Leskovec, Horvitz. Planetary-Scale Views on a Large Instant-Messenging Network. WWW 2008.

Social Networking Sites



My Facebook network. (Nexus)

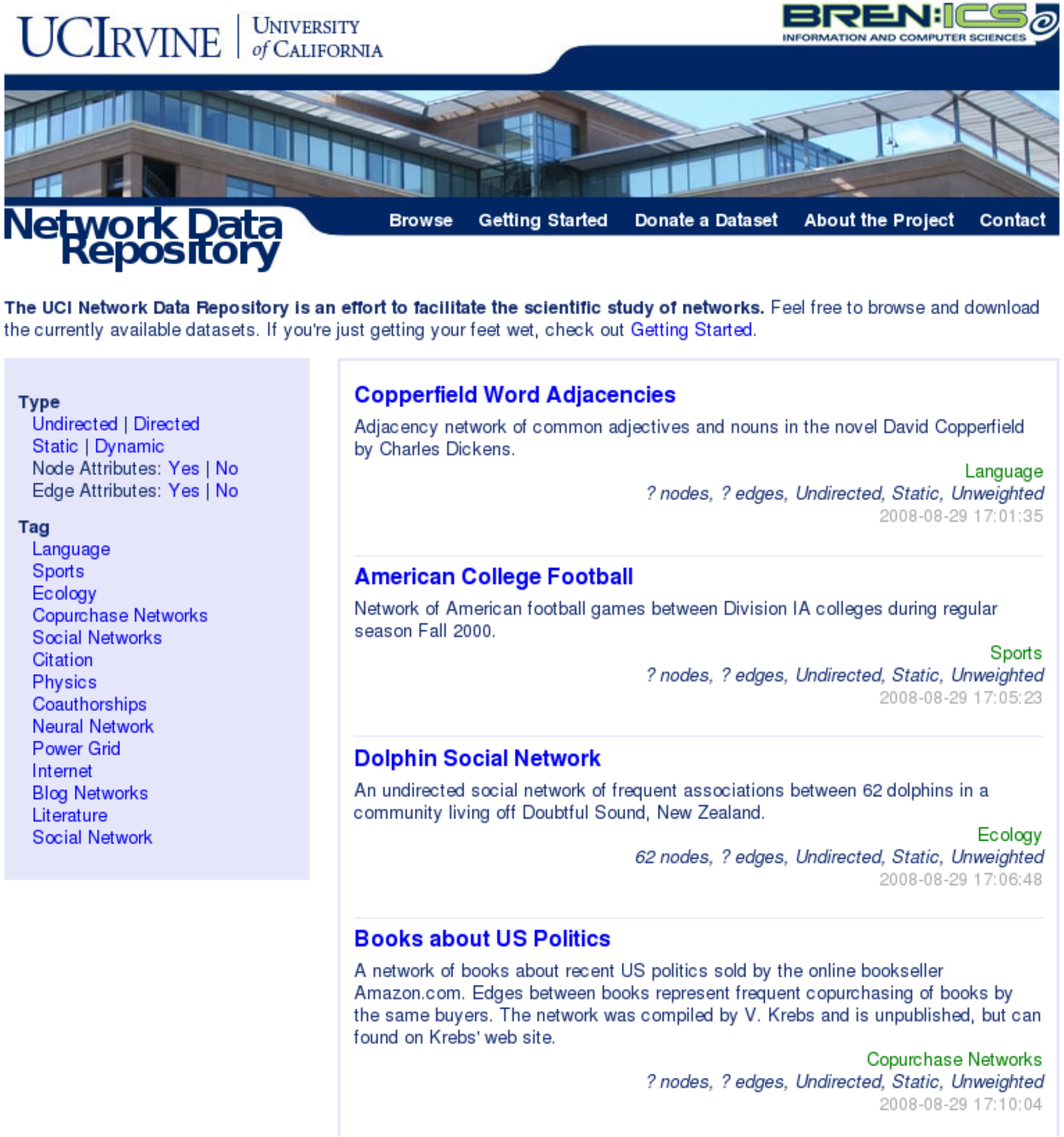
Wrapping up...

- Prevalence of data regarding online interaction
- Large populations, small timescales, many attributes

- Need for models and methods that can leverage these massive data sets for studying theories of large social systems

Open Access Network Data

- Browse, download network data
- Publicly available
- Facilitate network research



Network Data Repository

The UCI Network Data Repository is an effort to facilitate the scientific study of networks. Feel free to browse and download the currently available datasets. If you're just getting your feet wet, check out Getting Started.

Type
Undirected | Directed
Static | Dynamic
Node Attributes: Yes | No
Edge Attributes: Yes | No

Tag
Language
Sports
Ecology
Copurchase Networks
Social Networks
Citation
Physics
Coauthorships
Neural Network
Power Grid
Internet
Blog Networks
Literature
Social Network

Copperfield Word Adjacencies

Adjacency network of common adjectives and nouns in the novel David Copperfield by Charles Dickens.

Language
? nodes, ? edges, Undirected, Static, Unweighted
2008-08-29 17:01:35

American College Football

Network of American football games between Division IA colleges during regular season Fall 2000.

Sports
? nodes, ? edges, Undirected, Static, Unweighted
2008-08-29 17:05:23

Dolphin Social Network

An undirected social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand.

Ecology
62 nodes, ? edges, Undirected, Static, Unweighted
2008-08-29 17:06:48

Books about US Politics

A network of books about recent US politics sold by the online bookseller Amazon.com. Edges between books represent frequent copurchasing of books by the same buyers. The network was compiled by V. Krebs and is unpublished, but can found on Krebs' web site.

Copurchase Networks
? nodes, ? edges, Undirected, Static, Unweighted
2008-08-29 17:10:04