Hands-on Tutorial:
From Words to Networks:
Extraction and Analysis of
Semantic Network Data from Text Data

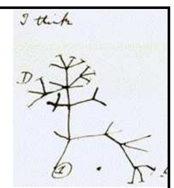
St. Petersburg State University, May 19, 2013
Jana Diesner, PhD
Assistant Professor
The iSchool @ University of Illinois
University of Illinois Urbana-Champaign

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

1

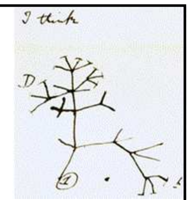
What we will do today



- Gain methodological and hands-on expertise in:
  1. **Information Extraction and Relation Extraction**
     - Distill relevant information from text data
     - Construct one-mode & multi-mode semantic networks from unstructured, natural language text data
     - Several natural language processing/ text mining techniques
       - Pre-processing
       - Identify salient concepts from single documents and corpora
       - Create and apply codebooks (aka dictionaries or thesauri)
       - Locate and classify entities that can serve as nodes for networks.
       - Link entities into edges (relation extraction)
  2. **Network Analysis**
     - Collect, visualize, analyze, interpret network data
       - Compute basic network metrics on the graph and node level

2

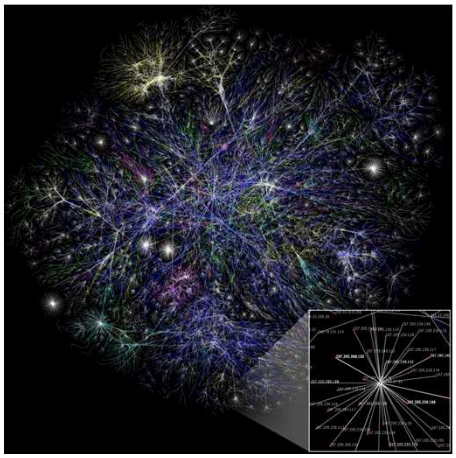
What we will do today



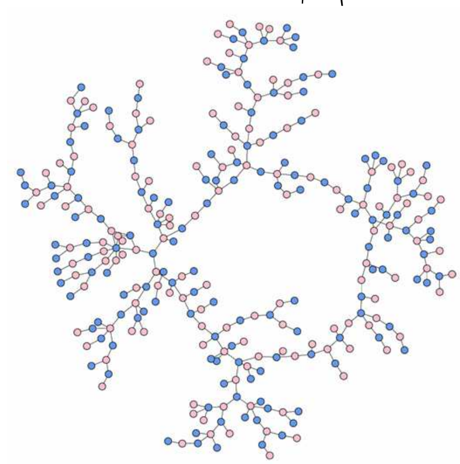
3. **Computational Thinking**
   - A fundamental skill that people from all backgrounds can use to solve problems in their domain.
   - Like reading, writing, arithmetic. Not a rote skill.
   - An approach to solving problems, designing systems and understanding human behavior that draws on concepts fundamental to computer science.

Introduction:
Network Analysis

The Concept of Networks



Socio-technical: The Internet



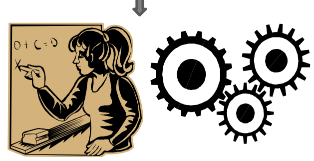
Social: High School Dating

Network Analysis: Workflow

Behavioral Data → **Data management and analysis** → **Utilization**

Interaction data

Data representation in relational form



Data analysis

- •Answer substantive and graph-theoretic questions
- •Develop and test hypothesis and theories
- •Create visualizations
- •Populate databases
- •Simulations, assess interventions
- •Input to further computations, e.g. machine learning

Questions that Network Analysis Helps to Answer

- General types of questions:
  - What does it look like? (Visualization)
  - What are the structure, functioning, dynamics of a network?
  - Who are the key entities? (Key player analysis)
  - Which subgroups exist? (Clustering)
  - What would happen if…? (Simulation)
- Specific questions:
  - Who talks to whom?
  - About what?
  - How do ideas and innovations emerge, spread, change and vanish in society?
  - Who are the key players in a network?
  - What benefits and risks result from an observed network structure for the network and its wider context?

Network Analysis: Core Idea and Relevance

- Concurrently study **nodes** (entities) and **edges** (relations)
- Understand **patterns of relationships** between entities
- Understand how **micro-level behavior** leads to **macro-level outcomes**

- Widespread acknowledgement of everything being connected
- Popularity of social networking services
- Advances in computational solutions for network analysis

Strong demand for solid knowledge & skills in network analysis in academia, administration, business.

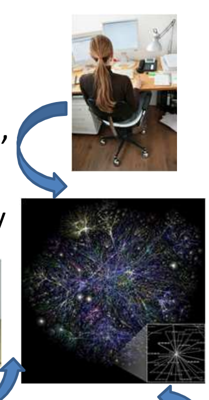
Network Basics: Nodes and Edges

- **Nodes** (aka points, vertices)
  - Classical: social agents (agents, groups): social network
  - One type of nodes: one-mode network
  - Multiple node classes (e.g. information and agents): multi-mode network
    - Socio-technical networks (people and infrastructures)
- **Edges** (aka links, ties)
  - Binary or weighted (frequency, probability, ordinal data)
  - Directed or not
  - Typed or not
    - One type: simplex, multiple types: multiplex

9

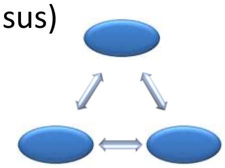
Network Basics: Network Characteristics

- **Structure**
  - Patterns of relations among entities
  - Network analysis: understand the structure, functioning and dynamics of networks
    - Consider entities & relations simultaneously
- **Dynamic**
- **Complex**
  - Multitude of interactions
  - Simple decision the node level can lead to complex structure and behavior



Network Basics: Levels of Analysis

- **Node:** Egocentric: ego and respective alters
- **Dyad:**
  – Undirected: (N square - N)/2
  – Directed: (N square – N)
  – Reciprocity?
- **Triad:**
  – Georg Simmel: triad smallest meaningful social unit
  – Directed: 16 isomorphic classes (triad census)
- **Cluster:** grouping
- **Complete network**



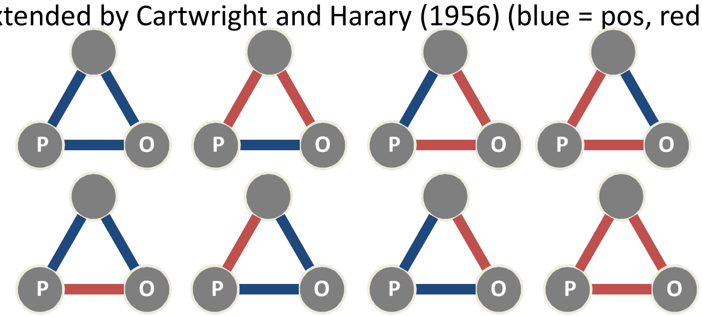
Levels of Analysis: Triads

- Exercise:
  – The enemy of my enemy is my friend
  – The friend of my friend is my friend
  – The friend of my enemy is my enemy
  – The enemy of my friend is my enemy

12

Network Basics: Structural Balance

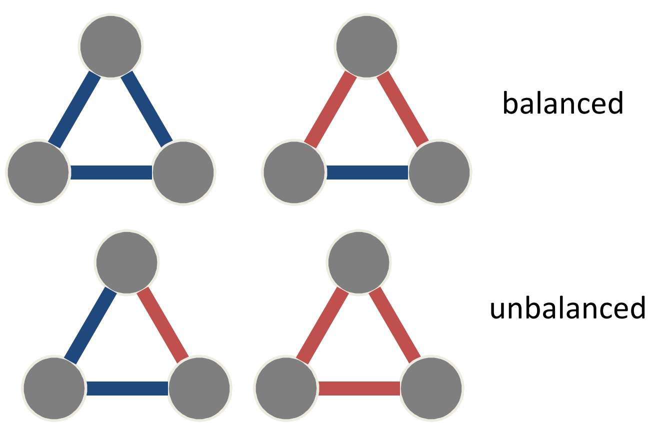
- Heider (1940): generalization of theory of cognitive dissonance, extended by Cartwright and Harary (1956) (blue = pos, red = neg)



- 1 or 3 +: balanced: No tensions, stable
- 0 or 2+: unbalanced: tension, stress, dissonance, change, unstable

Network Basics: Structural Balance

- A triad is balanced if its sign (product of signs) is positive



From Reading Tea Leaves to Metrics

- Who is key?
- Depends on: With respect to what dimension of power?

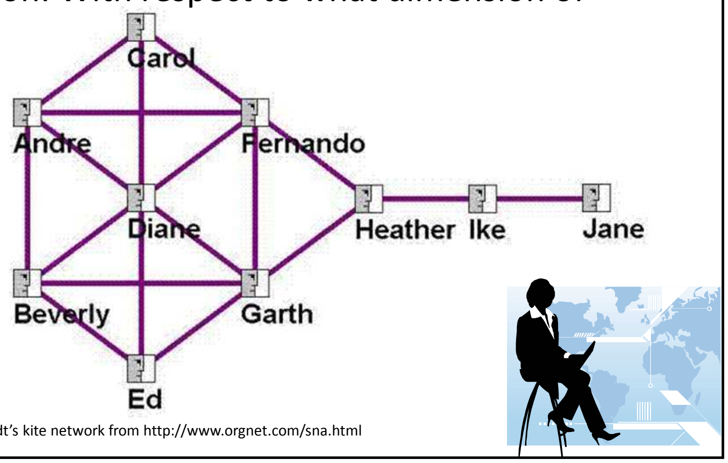


Image of David Krackhardt's kite network from http://www.orgnet.com/sna.html

Node-level Network Metrics: Dimensions of Power and Influence

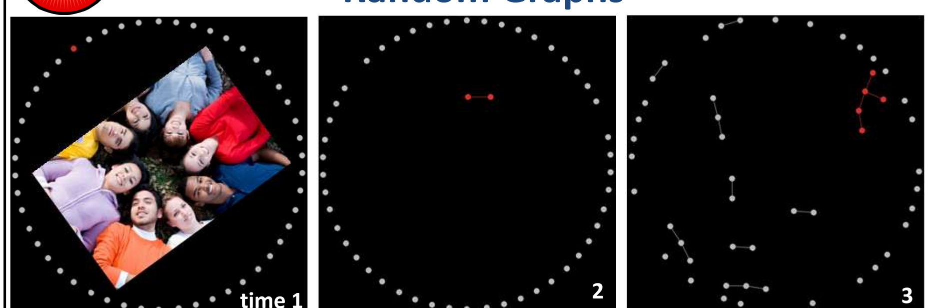
- Degree Centrality
  - Idea: immediate contacts (ego-network per node)
  - Power: Prestige, Action
  - Roles: Star, Hub
  - Computation: Sum of direct links per node
- Closeness Centrality
  - Idea: Reaching
  - Power: Fastest access to other nodes or what flows through the network
  - Roles: Monitor, Transmitter
  - Computation: Inverse of sum of geodesic distances (shortest path) from a node to all other nodes

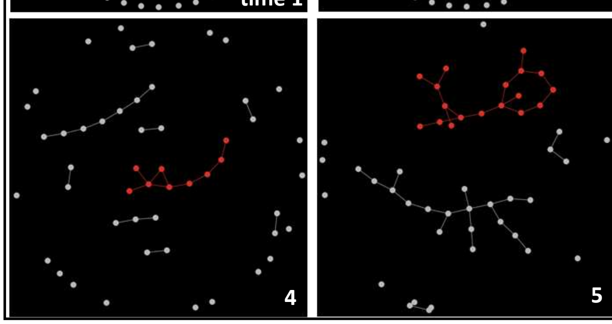
Node-level Network Metrics: Dimensions of Power and Influence

- Betweenness Centrality
  - Idea: Lying in between
  - Power: Control, Mediation
  - Roles: Broker, Gatekeeper, Bridge, Liason
  - Computation: Extent to which a node is on shortest path between any pair of nodes
- Eigenvector Centrality
  - Idea: Close to power
  - Power: Access
  - Roles: Lobbyist
- Equivalence
  - Regular (strict)
  - Structural (relaxed)
  - Idea: Redundancy, Resilience

Models of Network Evolution: Random Graphs

Run in NetLogo

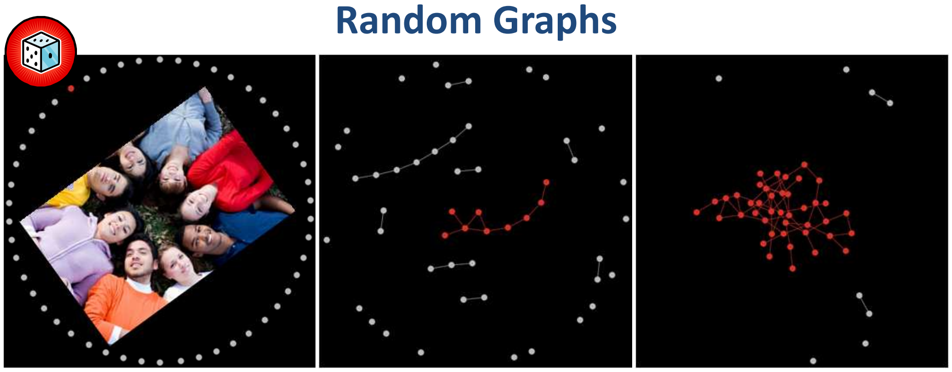




Principle: At each time step:
a) equal chance for any pair of unconnected nodes to get connected
b) one such pair picked randomly and linked

Question: What do you think will happen?

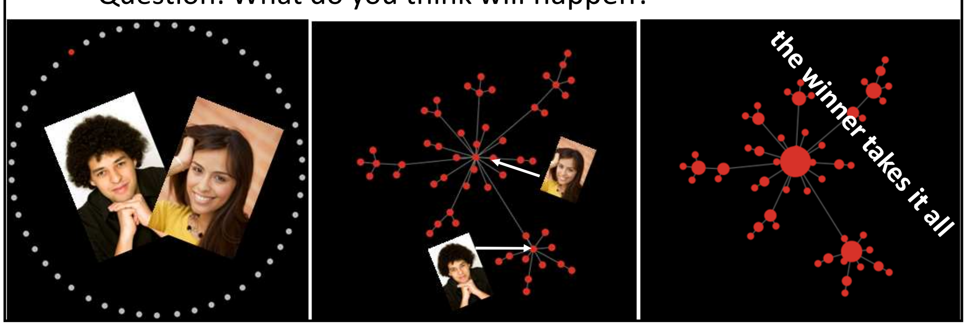
Models of Network Evolution: Random Graphs



- **Model: Random graphs (Erdős, Rényi 1959)**
- A giant component emerges
  - Component = connected group of nodes
- Node degree (degree centrality) in resulting network follows a normal distribution, no highly connected nodes

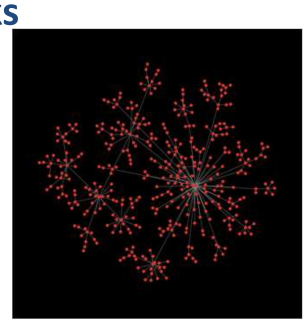
Models of Network Evolution: Scale-Free Networks

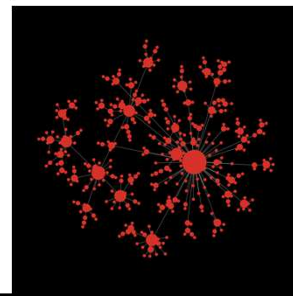
- Principle: Bias: a node's chance of getting linked is directly proportional to its degree.
- Example: Meet Ben and Amy. Everybody likes them. Therefore, at each time step, connections to Ben and Amy are a little more likely than all other connections.
- Question: What do you think will happen?



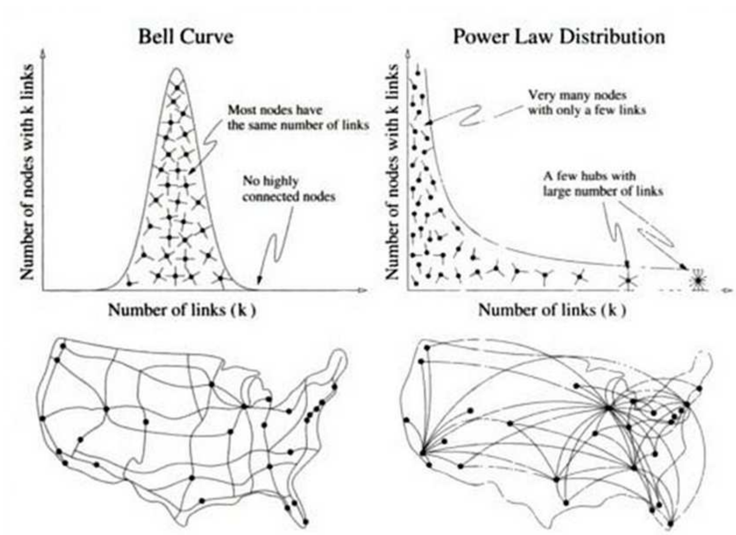
Models of Network Evolution: Scale-Free Networks

- **Model: preferential attachment,** aka **scale free networks, power-law networks (Barabási & Albert 1999)**
- Emergence of hubs (nodes with high degree centrality)
- A node's tenure and popularity translate into node's degree
  - This explains the first mover's advantage (e.g. Microsoft)
  - Question: How could Google as a late-comer in the search engine market win over the majority of users?
    - A: Fitness function





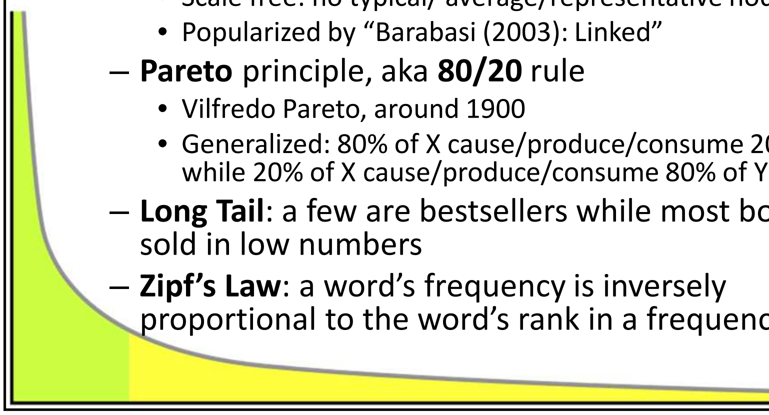
Random Networks vs. Scale-Free Networks



Barabasi (2002),pg. 71

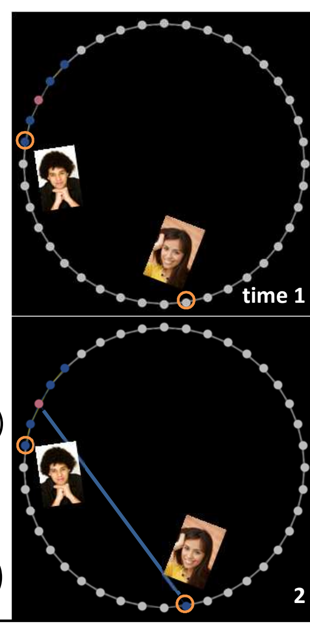
Scale-Free Networks

- Skewed distributions have several names/ flavors:
  - **Power law**: polynomial distribution with scale invariance
    - Scale free: no typical/ average/representative nodes
    - Popularized by “Barabasi (2003): Linked”
  - **Pareto** principle, aka **80/20** rule
    - Vilfredo Pareto, around 1900
    - Generalized: 80% of X cause/produce/consume 20% of Y, while 20% of X cause/produce/consume 80% of Y
  - **Long Tail**: a few are bestsellers while most books are sold in low numbers
  - **Zipf’s Law**: a word’s frequency is inversely proportional to the word’s rank in a frequency table



Models of Network Evolution: Small World

- You (red) know your friends (blue) and your friend’s friends (blue) (FOAF = friend of a friend)
- Principle: Rewiring (introduce randomness) Example:
  - Your friend’s friend Ben want to meet Amy.
    - How many people does he have to go through to be introduced to her?
    - Are you helpful in this process? (time 1)
  - Now lets assume you knew Amy… (t 2)
    - You = shortcut
    - Shortcuts make the world small
- **Model: Small World (Milgram 1967)**



Models of Network Evolution: Small World

- Milgram: experiment ($680 budget)

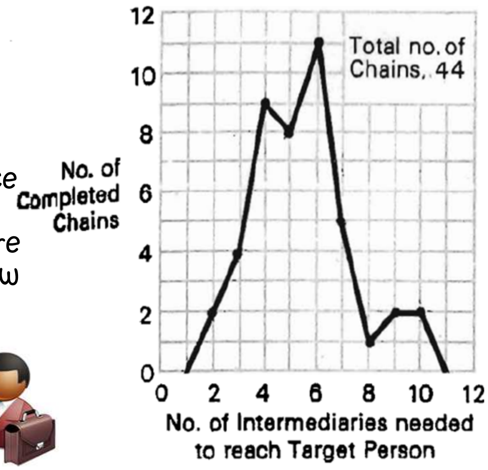
Wichita, KS

Omaha, NB

… send this to a personal acquaintance you know on a first name basis who is more likely than you to know the target person

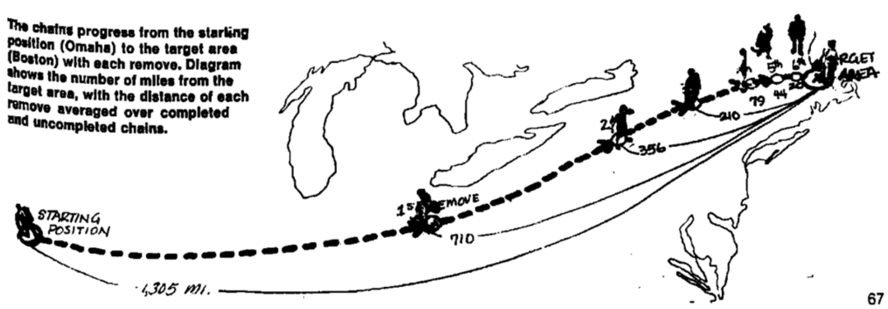
Wife of a divinity school teacher, Cambridge, MA

Stockbroker in Boston, MA



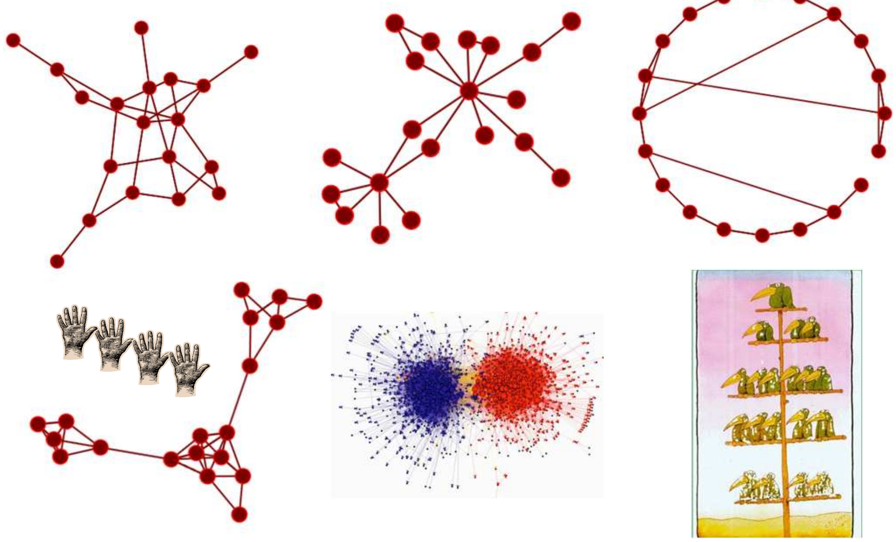
Models of Network Evolution: Small World

- Findings (Milgram, below): social distance > physical distance
- Replicated for Facebook (721 mio users, 69 billion links)
  - Average distance: 4.74, i.e. 3.74 intermediaries or "degrees of separation" (Backstrom et al. 2012)



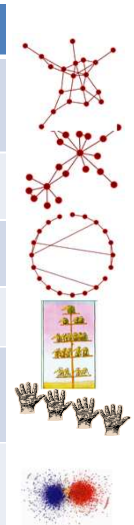
67

Network Topologies: Forms of Organization and Organizing



Network Topologies: Forms of Organization and Organizing

Name	Underlying Principle	Structural Fingerprint
Erdoes Renyi Random Graph	Randomness	Node degree follow normal distribution
Scale Free Network	Preferential Attachment	Most nodes have a few links while a few nodes have many links (hubs)
Small World Network	Shortcuts Friends know each other	Nodes connected to its neighbors and a few distant nodes
Hierarchy	Power	Directed, acyclic graph
Cellular Network	Balance between conceal and coordinate Trust	Dense connections within cells, sparse connections among cells.
Core Periphery Network	?	Nodes belong either to core or periphery. No ties between periphery nodes, but from core to core and periphery.



Differences to Other Domains

Statistics:
- General utility method
- Independence assumption
(iid: independent and identically-distributed random variables)

Social sciences:
- Reduction of social concepts and phenomena to unstructured data
- Focus on entities and their attributes
- Attributes static across social contexts
- Data collection via sampling

Bills with a bell curve on it is money from the past

http://de.wikipedia.org/wiki/Deutsche_Mark

Network Analysis:
- Becoming a general utility method
- Dependency assumption
- Express questions as structured variables
- Focus on entities and their relations
- Relations space and time dependent
- Some entity attributes can be formalized as relations
- Data collection via census

29

Network Basics: Network Analysis Process

1. Specification of goal, question, or task.
2. Specification of relevant nodes, edges, and network boundaries.
3. Data collection if no data given.
4. Representation of relational data as a list, matrix, or graph.
5. Analysis and utilization of relational data.
6. Result validation. Do error analysis if applicable.
7. Interpretation of the results with respect to step 1.

30

The network analysis process

- 1. Specify goal, question, or task
  - Identified as gap or contradiction in prior work
  - Given by client
- 2. Specify relevant nodes, edges, boundaries
  - Network boundaries:
    - Natural (all comments on a blog)
    - Demographic, ecological (all publications by UIUC researchers)

The network analysis process

- 3. Data collection
  - Complete population (all online information about GSLIS)
  - Snowballing
    - Problems?
      - Starting point
      - Missing isolates
  - Archival data
- 4. Representation of relational data as a list, matrix, or graph
  - All isomorphic
    - Express the same information without loss of information
    - Can be translated into one another (as opposed to transformed)

The network analysis process

- 5. Analysis and utilization of relational data
  - Data management: database operations: store, retrieve, search, merge
  - Heuristic: Visualization
  - Analytical: Network Analysis
  - Prognostic, exploring possible scenarios: Simulation
  - Input to machine learning system
  - Combination/ integration with classical statistics

Computing some metrics on network data is not the same as doing network analysis

The network analysis process

- 6. Result validation. Do error analysis if applicable.
  - Generalization
  - Limitations
- 7. Interpretation of the results with respect to step 1.
  - Implications
  - Suggestions, e.g. policies
  - Upper and lower bound for generalization: under what conditions do your findings hold truth?
  - Derive hypotheses and theories
  - Build models

Example: The network analysis process in action

Smartest guys in the room

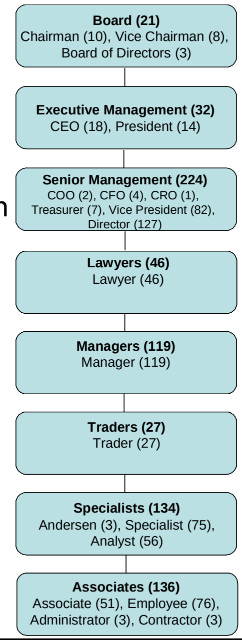
- 1985*: Kenneth Lay. Houston, Texas
- Business: gas supplier, energy broker, global commodity trader, and “other”
- $$ Success $$!
  - 2001: 7th-largest business organization (by revenue) in the USA, 21,000 employees in over 40 countries
  - Stock market’s darling
- 12/2001: Bankruptcy
- Charged with illicit accounting and business practices
- Involved Auditor: Arthur Andersen

Example: Network analysis process in action

1. Substantive, network related research questions:
   - How do the structure and functioning of an organization change during a crisis?
   - How does interpersonal communication change during a crisis?
2. Specification of relevant nodes, edges, network boundaries
   - Nodes: people, edges: email communication
   - Trade-off between coordination and concealment

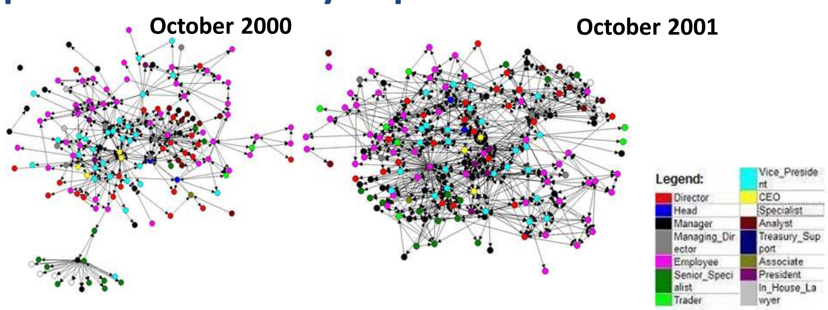
Diesner J, Frantz T, Carley KM (2005) Communication Networks from the Enron Email Corpus "It's Always About the People Enron is no Different”. Computational and Mathematical Organization Theory (CMOT), 11(3), 201-228.

Example: Network analysis process in action

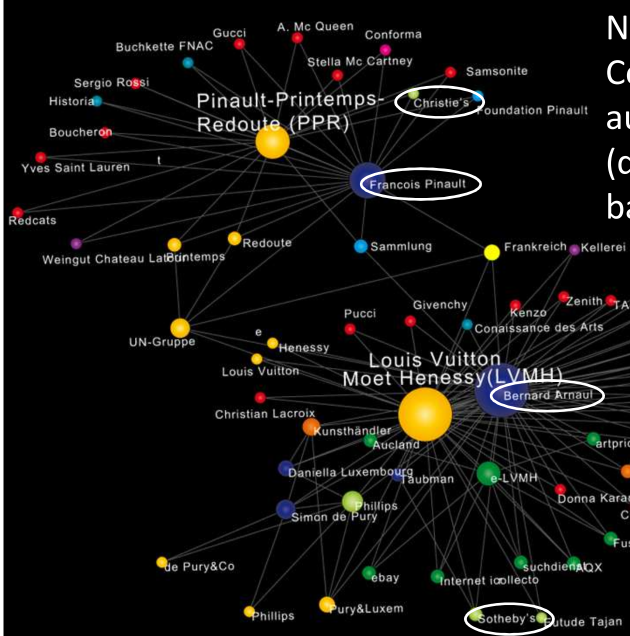


3. Data:
- Inquiries by SEC and FERC in 2002
  -> release of 620,000 emails
- Real communication from a real organization (over time 10/1998 - 07/2002)
- Rare glimpse into organizational processes, culture, crisis
- from **raw relational data** (email addresses) to **network data** (people, content, context) to **insights and knowledge**
- Compiled career history for 676 people with full name, email addresses (1 to 17 per person, average: 2.2), career history (110 job titles, 13 positions, 8 ranks)
- Social structure (headers)
- Semantics (bodies)

Example: Network analysis process in action



5., 7. Network Analysis and Interpretation -> Answers:
- Network: higher density, heterogeneity, cohesion, new links
- More by-passing of formal chains or hierarchies
- Pair-wise relations intensified (trusted others), triads decreased
- More communication, but communication diversity and entropy decrease (less words, less distinct words, higher redundancy)
  - Discourse drifts towards polarized ends of themes and issues?
- Individual patterns start to strongly vary depending on addressee



Network Visualization: Co-ownership beyond auction houses (qualitative, text data based)

Network Basics: Network Visualizations

- Heuristic utility
- Appropriate for:
  - Stimulating communication
- Inappropriate for:
  - Large-scale data
  - Objective analysis
  - Be careful when used for consulting
- Layout options
  - Idea of spring embedders
  - Diplomatic: Circles

Bravo! You have passed the primer in network analysis!

41

Readings and References

- Network Analysis – introductory text books:

  - Hanneman, RA & Riddle, M. (2005). Introduction to social network methods. Riverside, CA: University of California. URL: http://www.faculty.ucr.edu/~hanneman/nettext/
  - Wasserman, S. and K. Faust, Social Network Analysis: Methods and Applications Cambridge University Press), 1994.
  - Easley, D. & Kleinberg, J. (2010). Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press. URL: http://www.cs.cornell.edu/home/kleinber/networks-book/

- Network models – we ran them in NetLogo:

  - Erdős, P., & Rényi, A. (1959). On random graphs. Publicationes Mathematicae Debrecen, 6, 290-297.
  - Barabási, A., & Albert, R. (1999). Emergence of Scaling in Random Networks. Science, 286(5439), 509.
  - Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. Nature, 393, 440-442.
  - Milgram, Stanley. (1967). The Small World Problem, Psychology Today, 2: 60-67.
  - Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2011). Four degrees of separation. Arxiv preprint arXiv:1111.4570.

42

Readings and References

- On skewed distributions:
  - Anderson, Chris (2006). The Long Tail: Why the Future of Business Is Selling Less of More. New York: Hyperion.
  - Barabasi, A.L. (2002). Linked: the new science of networks Perseus Publishing, Cambridge.
  - Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. Contemporary Physics 46: 323–351. doi:10.1080/00107510500052444.
  - Simon, H. A. (1955). On a Class of Skew Distribution Functions. Biometrika 42: 425–440. doi:10.2307/2333389.
  - Zipf, George K. (1949) Human Behavior and the Principle of Least-Effort. Addison-Wesley.
- Triads: Simmel, G. (1950). The Sociology of Georg Simmel. Free Press.
- Balance Theory: Heider, F. (1982). The psychology of interpersonal relations: Lawrence Erlbaum.
- Computational Thinking: Wing, J. M. (2006). Computational thinking. CACM, 49(3), 33 - 35.

References for Images

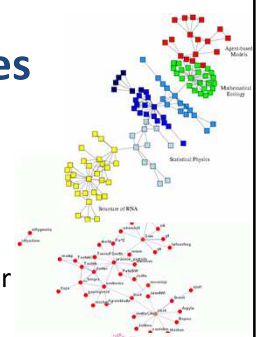
- Internet: http://en.wikipedia.org/wiki/File:Internet_map_1024.jpg
- High school dating: Data from P.S. Bearman, J. Moody, & K. Stovel, Chains of affection: The structure of adolescent romantic and sexual networks, American Journal of Sociology 110, 44-91 (2004), image by Mark Newman http://www-personal.umich.edu/~mejn/networks/
- Art markets: Diesner J., Stützer, C. (2008) Finding relations. Presentation at Chemnitz Art Museum.
- Core periphery: Image from http://wwwpersonal.umich.edu/ ladamic/img/politicalblogs.jpg)
- Skewed distribution: http://en.wikipedia.org/wiki/File:Long_tail.svg
- Enron: http://www.pbs.org/independentlens/enron/index.html

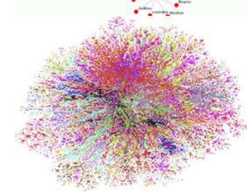
Network Analysis: Learning Resources

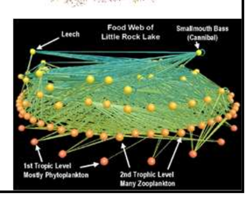
- Mailing list:
  - http://www.insna.org/pubs/socnet.html
- Organization:
  - International Network for Social Network Analysis (INSNA): http://www.insna.org/index.html
- Journals:
  - Social Network Analysis and Mining http://www.springer.com/computer/database+management+%26+information+retrieval/journal/13278
  - Network Science: http://www.indiana.edu/~netsci/index.html
  - Connections: http://www.insna.org/pubs/connections/index.html
  - Social Networks: http://www.sciencedirect.com/science/journal/03788733

Network Data Sets: Types

- Social
  - From small-scale (observations) to web-scale (social media)
- Collaboration (who works with whom)
  - Co-author, co-appear in movies, co-edit, co-chair (corporate boards)
- Communication
  - Who talks to whom
- Information
  - The web, citation networks, semantic networks
- Technological
  - Routers, power generation stations
- Biological
  - Food web, animals, cell metabolism, neurological







For typology see Kleinberg textbook, chapter 2.4, Images http://www-personal.umich.edu/~mejn/networks/

Network Data Sets

- Classic small scale, small group data: http://vlado.fmf.uni-lj.si/pub/networks/data/ucinet/ucidata.htm (also at https://sites.google.com/site/ucinetsoftware/datasets)
- Smaller collection of classic and some older internet datasets: http://www-personal.umich.edu/~mejn/netdata/
- Large scale, social media data: http://snap.stanford.edu/data/index.html
- More internet datasets: http://law.di.unimi.it/datasets.php
- (Socio)Technical networks: http://www.sommer.jp/graphs/

Network Data Sets

- Benchmark datasets and competitions: http://hcil.cs.umd.edu/localphp/hcil/vast/archive/viewbm.php
- Kitchen sink sorted by type: http://networkdata.ics.uci.edu/index.html
- Large kitchen sink: http://www.casos.cs.cmu.edu/computational_tools/data2.php
- A smaller kitchen sink: https://nwb.slis.indiana.edu/community/?n=Datasets.HomePage
- And an even smaller kitchen sink: http://pajek.imfm.si/doku.php?id=data:pajek:vlado

Network Data Sets: Collection

- NodeXL (http://nodexl.codeplex.com)
- Netscrape (http://socialcomputing.asu.edu/pages/netscrape)
- Codebase: ScraperWiki https://scraperwiki.com/

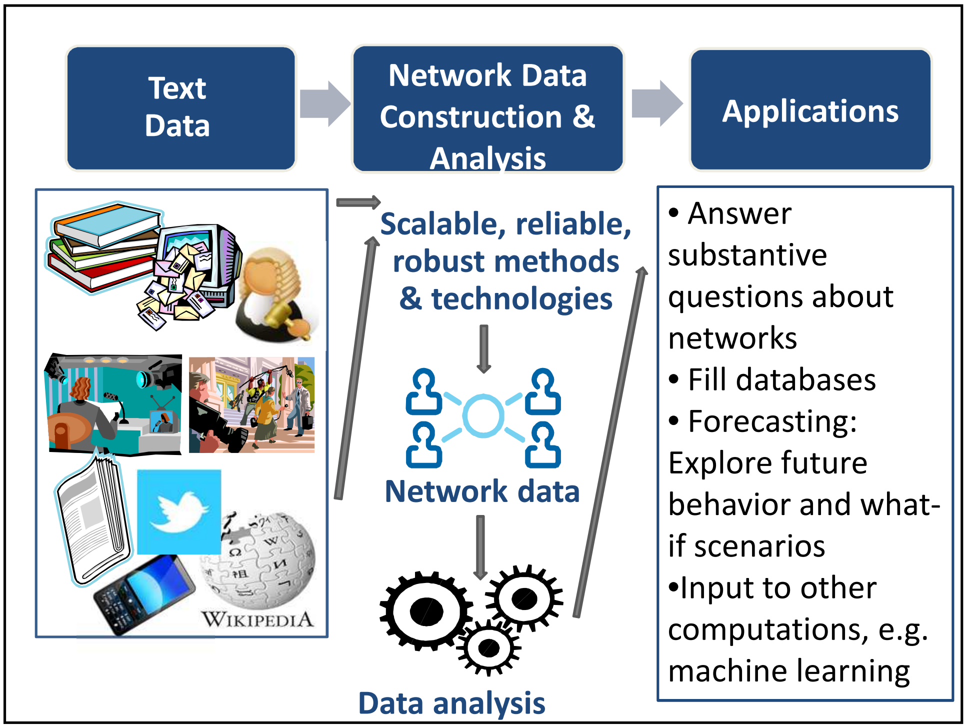
Social Network Analysis Software: Overview and Main Stream Tools

- Overview: http://en.wikipedia.org/wiki/Social_network_analysis_software (Name and URL, main functionality, input and output formats, platforms, license, costs)
- Mature in terms of metrics, matrix transformation routines, visualization, import and export from/ to various data formats
  - Biggest player: UCINET (https://sites.google.com/site/ucinetsoftware/home)
    - limited scalability, commercial, free trial
  - Pajek (http://pajek.imfm.si/doku.php)
    - Comparatively great scalability, free
  - visone: http://visone.info/ (free for academic research)
  - NetMiner (http://www.netminer.com/index.php) (commercial)

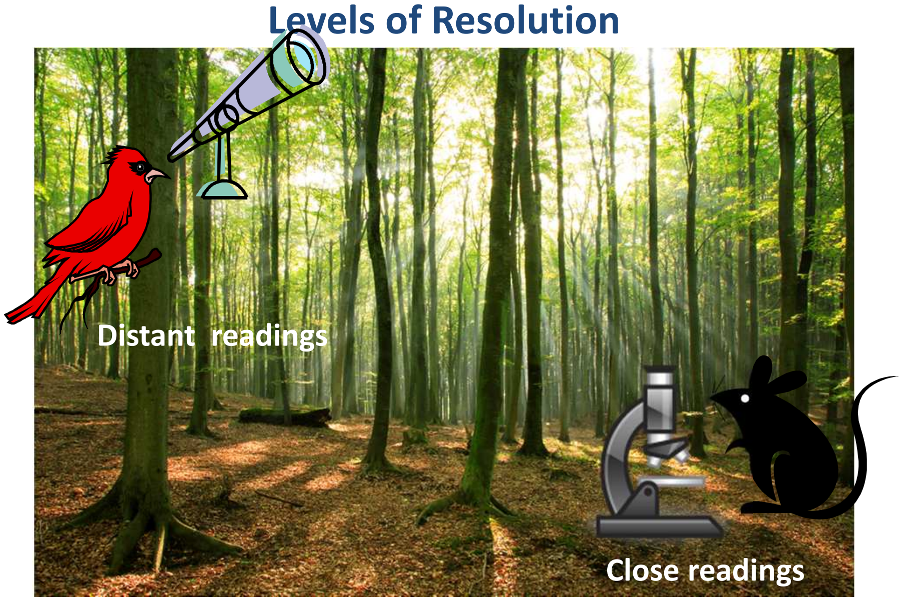
Network Analysis Software: Open Source

- Low entry, high ceiling: NodeXL (http://nodexl.codeplex.com/) (grouping, viz, baseline set of metrics)
- Currently popular: Gephi: http://gephi.org/ (viz, format conversion)
- R routines/ packages for SNA (http://erzuli.ss.uci.edu/R.stuff/), now mainly integrated into statnet (http://www.statnet.org/)
- Highly flexible graph manipulation code base: JUNG http://jung.sourceforge.net/ (no update since Jan 2010)
- Library and GUI-based tool: GUESS (uses JUNG) (http://graphexploration.cond.org/download.html#source) (no update since 2007)

From Words to Networks: Methodology

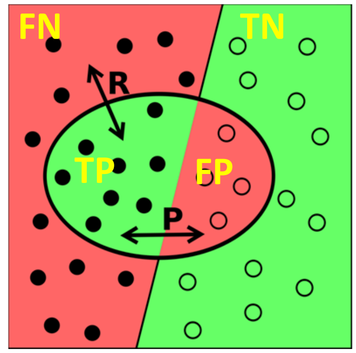


Text Analysis and Social Science and Computing: Levels of Resolution



Accuracy Assessment in Information Retrieval: Concepts of recall and precision

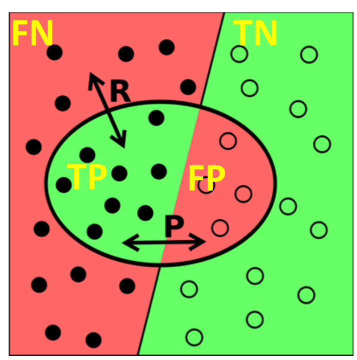
- Relevant items: left of straight line
- Retrieved items: in oval
- Red regions: errors:
  - Left: false negatives
  - Right: false positives
- Precision P:
  left green region/ oval
- Recall R:
  left green region/ left region



	Truth	
Result	TP: True positive (yeah, correct result)	FP: False positive (oops, false alarm)
	FN: False negative (oops, blind spots)	TP: True negative (yeah, correctly missing results)

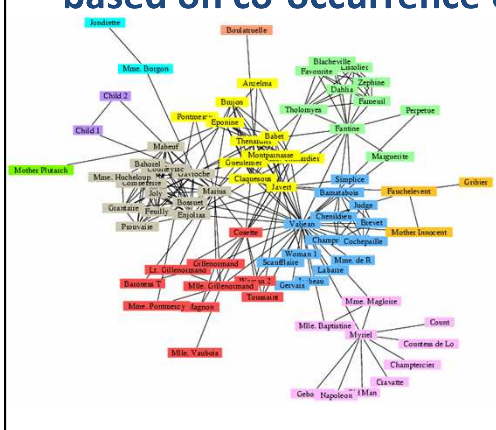
Accuracy Assessment in Information Retrieval: Concepts of recall and precision

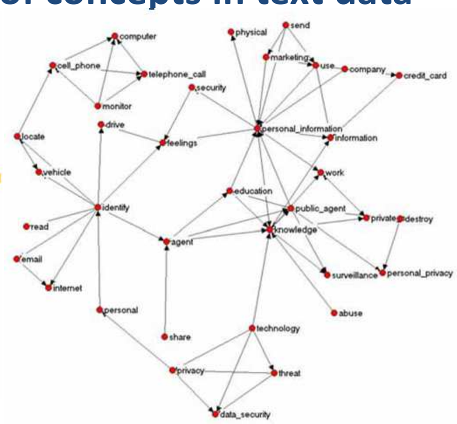
- Recall = TP/ TP+FN
- Precision = TP/ TP+FP
- Relationship?
  - Inverse. Thus a harmonic mean is calculated:
  - F= T*P / 0.5 (T+R)



Text Analysis and Social Science and Computing: Levels of Resolution

From Words to Networks: Association networks, based on co-occurrence of concepts in text data





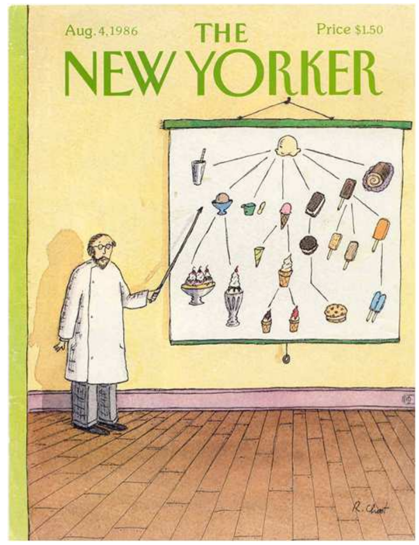
- Basic principle and assumption: Language, information and knowledge can be modeled and represented as relational data

Les Miserables: M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Physical Review E 69, 026113 (2004).

Diesner J, Kumaraguru P, Carley KM (2005) Mental Models of Data Privacy and Security Extracted from Interviews with Indians. 55th Annual Conference of International Communication Association (ICA), New York, NY, May 2005.

Classification: A main Task in Text Coding

- Ontology: the study of being or existence
- Taxonomy: practice & science of classification
- Modifying ontologies:
  - Human updating, re-indexing
- Solutions?
  - Citizen science, crowd-sourcing
  - Automation

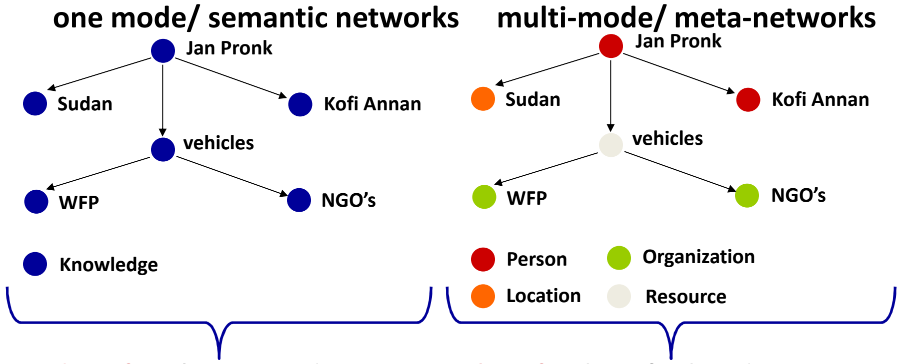


Relation Extraction: One-mode and Multi-mode Networks

Example from UN News Service (New York), 12-28-2004: "Jan Pronk, the Special Representative of Secretary-General Kofi Annan to Sudan, today called for the immediate return of the vehicles to World Food Programme (WFP) and NGOs."

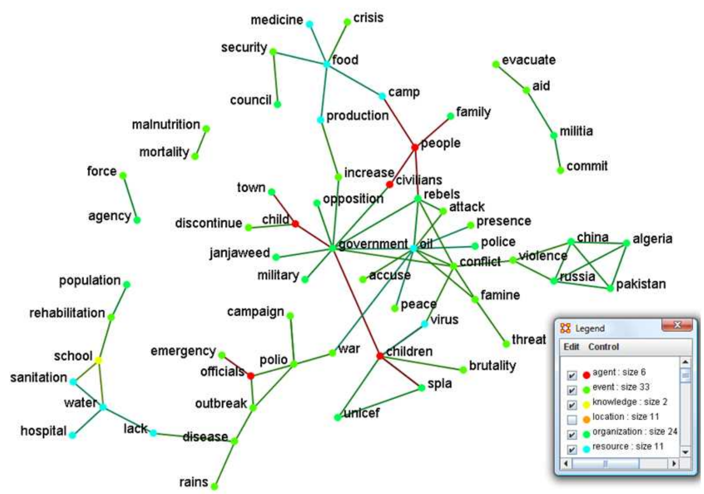
extract relational data:

one mode/ semantic networks **multi-mode/ meta-networks**



Identify relevant nodes **Classify identified** nodes

From Words to Networks: Relation Extraction



Diesner, J., & Carley, K. M. (2005). Revealing Social Structure from Texts: Meta-Matrix Text Analysis as a novel method for Network Text Analysis. In V. K. Narayanan & D. J. Armstrong (Eds.), Causal Mapping for Information Systems and Technology Research (pp. 81-108). Harrisburg, PA: Idea Group Publishing.

Motivation for Relational Text Analysis

- **Fact:** Collection and storage of large volumes of text data cheap, easy and efficient
  - Interviews, books, news wire articles, legal documents, annual reports, data from web 1.0 (web sites) and web 2.0 (emails, blogs, chats, …)
- **Need**: Methods and tools for automated, robust and reliable knowledge discovery and reasoning about information, incl. network structures, from text data.
- **Challenge:** Effective, efficient and controlled extraction of relevant (user-defined) instances of categories (e.g. node and edge classes) from unstructured, natural language text data.

Basic Types of Information in Text Data

- **Morphology**: structure of words
  - E.g. spelling, inflections, derivations
- **Syntax**: relationships between words
  - e.g. parts of speech tagging
- **Semantics**: meaning of language
  - e.g. word sense disambiguation, grammars
- **Pragmatics**: language in context and social use of language
  - e.g. sentiment analysis, discourse analysis
- **Relation Extraction (this lab)**: borrows from all of the above

Network Data Extracted from Texts

- Respective theories and methods developed across many disciplines:
  - Artificial Intelligence (e.g. Sowa)
  - Cognition and Linguistics (e.g. Collins)
  - Communications (e.g. Doerfel, Monge)
  - Political Science (e.g. Schrodt)
  - Sociology (e.g. Carley, Mohr)
  - Computer Science (e.g. McCallum)

Methods for Constructing Networks of Words

Method	Automation	Abstraction	Generalization
1. Mental Models (Spreading Activation) (Collins & Loftus 1975)			
2. Case Grammar and Frame Semantics (Fillmore 1982, 1986)			
3. Discourse Representation Theory (Kamp 1981)			
4. Knowledge representation in AI, assertional semantic networks (Shapiro 1971, Woods 1975)			
5. Centering Resonance Analysis (Corman et al. 2002)			
6. Mind maps (Buzan 1974)			
7. Concept maps (Novak & Gowin 1984)			
8. Hypertext (Trigg & Weiser 1986)			
9. Qualitative text coding (Grounded Theory) (Glaser & Strauss 1967)			
10. Definitional semantic networks incl. text coding with ontologies (Fellbaum 1998)			
11. Semantic Web (Berners-Lee et al. 2001, Van Atteveldt 2008)			
12. Frames (Minsky 1974)			
13. Semantic Grammars (Franzosi 1989, Roberts 1997)			
14. Network Text Analysis in social science (Carley & Palmquist 1991)			
15. Event Coding in pol. science (King & Lowe 2003, Schrodt et al. 2008)			
16. Semantic networks in comm. science (Danowski 1993, Doerfel 1998)			
17. Probabilistic graphical models (Howard 1989, Pearl 1988)			

Diesner, J. (2012). Uncovering and managing the impact of methodological choices for the computational construction of socio-technical networks from texts. CMU-ISR-12-101, Carnegie Mellon University.

Exercise

- Task: Represent the relevant information contained in the text data as network data.
- Questions:
  - What criteria did you use to identify nodes? Edges?
  - How did you come up with your criteria?
  - How many different criteria (features) did you use?
  - How consistent were you in applying your criteria?
  - How similar are the solutions from different teams?

Lab:

- **Information Extraction**
- **Relation Extraction**

Lab: Three Step Process

- **Text pre-processing**:
  - Natural Language Processing (NLP) techniques precondition for finding meaningful information, incl. representations of nodes and edges, in text data
  - Selection of relevant entities (positive filter: thesaurus, entity extraction) or removal of irrelevant entities (negative filter: delete list) given the research question, data, domain
- **Node identification (and classification)**
  - Thesaurus-based (manually or automatically built)
- **Edge identification (and classification)**
  - Identification: Proximity based approach (co-occurrence)

Finding salient terms: Cumulative frequency

- Bag of Words
- How to:
  - Load texts into AutoMap
  - Create Union Concept List
    - Generate: Concept List: Union
  - Sort results file by decreasing cumulative frequency
  - This is one dimension of salience, prominence, importance
  - What are other dimensions?

Zipf, George Kingsley (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, Mass.: Addison-Wesley

Refine Bag of Words: Stop words

- Stop words listed in delete list
- Serves as negative filter (remove from text data what's contained in list)
- How to:
  - Preprocess: Text Refinement: Apply delete list
  - How to construct a delete list?
    - In concept list viewer
    - Use predefined lists for English from "apply delete list panel"
    - Construct your own, one entry per line (incl. n-grams), save as .csv file
  - Notion of adjacency (direct vs. rhetorical, which maintains original distance of words)

Finding salient terms: TD*IDF

- What determines word's importance in corpus?
  - Discriminating and distinguishing
- tf = term frequency (importance of term within document)

$$tf = \frac{\textit{cumulative occurrence of term x in document y}}{\textit{total number of terms in document y}}$$

- idf = inverse document freq. (importance of a term in corpus)

$$idf = \log \frac{\textit{total number of documents in corpus}}{\textit{total number of documents containing term x}}$$

*tfidf = tf * idf*

- tfidf: strategy and measure
  - High if tf = high and df = low
  - High for signal, low for noise
- How to: Generate: Concept List: Union

Finding salient terms: N-grams

- Meaningful multi-words units
- How to:
  - Generate: generalization thesaurus: bigram
  - Sort by decreasing frequency and decreasing tfidf, pick suitable entries, removed duplicates

Text pre-processing: Stemming

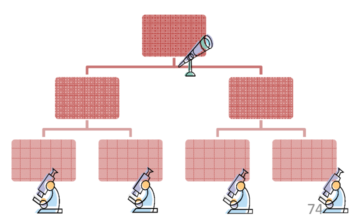
- Detects inflections and derivations of concepts
- Converts each term into its morpheme
- How to:
  - Pre-process: text refinement: stemming
- Two families of stemmers:
  - Porter (rule-based): high efficiency, poor human readability
  - Krovetz (dictionary-based): lower efficiency, better human readability

Porter, M.F. 1980. An algorithm for suffix stripping. *I 14* (3): 130-137.

Krovetz, Robert (1995). *Word Sense Disambiguation for Large Text Databases*. Unpublished PhD Thesis, University of Massachusetts.

Text pre-processing: When to stop?

- When to stop? (“criteria”)
  - Orcam’s razor:
    - 14th-century English logician and Franciscan friar, William of Ockham.
    - Aka lex parsimoniae (law of parsimony)
    - Basic idea: All other things being equal, the simplest solution is the best.
    - Why does it matter?
      - Don’t want: Overfitting
      - Want: Generalizability



Positive filter: Thesaurus

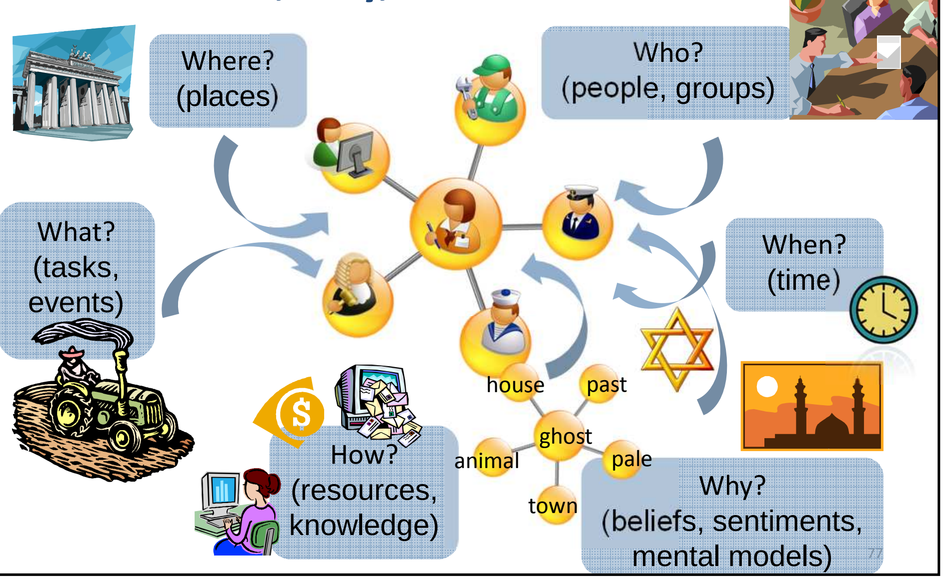
- Text term (incl. n-gram), concept, entity class
- Functions: disambiguation (1,2), consolidation (3,4), n-gram concatenation (3,4)
- Examples:
  1. Apple, apple, organization
  2. apple, apple, resource
  3. Digital Humanities, digital_humanities, knowledge
  4. computational folkloristic, digital_humanities, knowledge

Positive filter: Thesaurus Construction

- How to use thesaurus:
  - List relevant entities: serves as positive filter
    - Then construct one-mode network, aka semantic network
  - Cross-classify relevant entities with ontological categories
    - Then construct multi-mode network
- Help for constructing thesauri:
    - Computer-supported: union Concept List (terms with highest frequency and tfidf values after deletion), bigrams
    - Automated: AutoMap: generate: thesaurus suggestion (more on slide 42)
    - External sources (e.g. CIA World Fact Book, WordNet)
    - Other automated techniques, e.g. Bootstrapping
- Limitations:
  - Tedious, incomplete, outdated, deterministic

76

Ontology: the who, what, when, where, why, and how of events



Linking nodes: Approaches

- **Syntax and surface patterns** (Fillmore, Schrodt)
  - Linguistics: parsing trees
- **Logical and Knowledge Representation in Artificial Intelligence** (Shapiro)
  - first order calculus, predicate logic (quantifiers)
- **Distance based** (Danowski)
  - Communications: distance in text (windowing) or in space (Euclidean)
- **Probabilistic, learning from data** (McCallum)
  - Machine learning techniques: probabilistic (Bayesian), kernels (N-dimensional similarity), graphical models (hidden markov models, conditional random fields), boot strapping

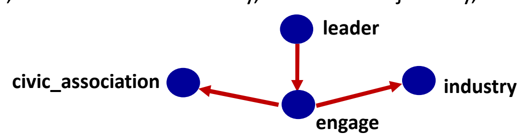
Cites and summary in Diesner, J., & Carley, K. M. (2010). Relation Extraction from Texts (in German, title: Extraktion relationaler Daten aus Texten). In C. Stegbauer & R. Häußling (Eds.), Handbook Network Research (Handbuch Netzwerkforschung) (pp. 507-521). Vs Verlag.

Link formation: one simple approach: Distance-based

- Distance based approach, Windowing:
  - Text Unit (text, paragraph, sentence)
  - Window Size (2 to N)
  - Adjacency (direct or rhetorical)

Leader	xxx	actively	involved	xxx	several	industry	xxx	civic	associations.

- Exercise: given the following thesaurus, what combination of distance based features results in a useful relational structure?
  - Thesaurus: leader, leader; involved, engage; industry, industry; civic associations, civic_association
- Sentence, thesaurus content only, rhetorical adjacency, window size 5:



Danowski, J. (1982). A network-based content analysis methodology for computer-mediated communication: An illustration with a computer bulletin board. In R. Bostrom (Ed.), *Communication Yearbook*, 6: 904-925. New Brunswick, NJ: Transaction Books.

Distance-based node linkage

- How to:
  - AutoMap: Generate: meta network: meta-network dynetlml (Union)
  - Set parameters – for this course, mainly select:
    - Directionality: bidirectional
    - An appropriate window size
    - An appropriate stop unit
    - A meta-network thesaurus, and check the box: use master thesaurus format
      - » The thesaurus needs to have the following header row:
      - » conceptFrom, conceptTo, metaOntology, metaName
      - » You don't need a value in the last column

Thesaurus Construction

- From rule-based and deterministic methods to probabilistic and machine-learning based methods
- How to use it in AutoMap:
  – Load raw data, no preprocessing
  – Generate: thesaurus suggestion
    - Decision support wizard on that panel has overview on types
    - Most accurate one: the model in the middle, sufficient for most work in this course: the one right above the one in the middle

References

- Recommended further readings related to this session:

  – McCallum, A. (2005). Information extraction: distilling structured data from unstructured text. ACM Queue, 3(9), 48-57.
  – Diesner, J., Carley, K. M. (2011): Semantic Networks. In G. Barnett (Ed), Encyclopedia of Social Networking, (pp. 595-598). Sage Publications.
  – Diesner, J., Carley, K. M. (2011): Words and Networks. In G. Barnett (Ed.), Encyclopedia of Social Networking, (pp. 958-961). Sage Publications.

References

- Image for Precision and Recall from Wikipedia
- Mental Models:
  - Johnson-Laird, P. (1983). *Mental Models*. Cambridge, MA: Harvard University Press.
  - Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management*, 20, 403-437.
  - Rouse, W. B., & Morris, N. M. (1986). On looking into the black box; prospects and limits in the search for mental models. *Psychological Bulletin*, 100, 349-363.

Readings and References

- Baker, W. E., & Faulkner, R. R. (1993). The Social Organization of Conspiracy: Illegal Networks in the Heavy Electrical Equipment Industry. *American Sociological Review, 58*(6), 837-860.
- Carley, K. M. (1997). Network text analysis: The network position of concepts. In C. W. Roberts (Ed.), *Text Analysis for the Social Sciences: Methods for Drawing Statistical Inferences from Texts and Transcripts* (pp. 79–100). Mahwah, NJ, USA: Lawrence Erlbaum Associates, Inc.
- Diesner, J., & Carley, K. M. (2010). Relational Methods in Crime Research and Law Enforcement. In C. Stegbauer & R. Haeussling (Eds.), *Handbook Network Research*: Vs Verlag.
- Diesner, J., Frantz, T. L., & Carley, K. M. (2005). Communication Networks from the Enron Email Corpus. "It's Always About the People. Enron is no Different". *Computational & Mathematical Organization Theory, 11*(3), 201-228.
- Doerfel, M. (1998). What Constitutes Semantic Network Analysis? A Comparison of Research and Methodologies. *Connections, 21*(2), 16-26.
- Mergel, I., Diesner, J., & Carley, K. M. (2010). *Attention Networks among Members of Congress*. Paper presented at the XXX International Sunbelt Social Network Conference.
- Mohr, J. (1998). Measuring Meaning Structures. *Annual Reviews in Sociology, 24*(1), 345-370.
- Monge, P. R., & Contractor, N. (2003). *Theories of Communication Networks*. New York: Oxford University Press.
- Schrodt, P. A., & Gerner, D. J. (1994). Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982-92. *American Journal of Political Science, 38*(3), 825-854.
- Sowa, J. (1992). Semantic Networks. In S. C. Shapiro (Ed.), *Encyclopedia of Artificial Intelligence* (2nd ed., pp. 1493 - 1511). New York, NY, USA: Wiley and Sons.

Acknowledgements

- This work was supported by the National Science Foundation (NSF) IGERT 9972762, the Army Research Institute (ARI) W91WAW07C0063, the Army Research Laboratory (ARL/CTA) DAAD19-01- 2-0009, the Air Force Office of Scientific Research (AFOSR) MURI FA9550-05-1-0388, the Office of Naval Research (ONR) MURI N00014-08-11186, and a Siebel Scholarship. Additional support was provided by CASOS, the Center for Computational Analysis of Social and Organizational Systems at Carnegie Mellon University. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, ARI, ARL, AFOSR, ONR, or the United States Government.
- I thank Nikita Basov and the Centre for German and European Studies at the St. Petersburg State University, Russia, for hosting this workshop.

Thank you!
Q&A

- For any questions, comments, feedback, follow-up- now and in the future:

Jana Diesner
Email: jdiesner@illinois.edu
Phone: ++1 (412) 519 7576
Web: http://people.lis.illinois.edu/~jdiesner