ECML 2007 PKDD
WARSAW POLAND

THE 18TH EUROPEAN CONFERENCE ON MACHINE LEARNING
AND
THE 11TH EUROPEAN CONFERENCE ON PRINCIPLES AND PRACTICE OF KNOWLEDGE DISCOVERY IN DATABASES

DISCOVERING AND TRACKING USER COMMUNITIES

TUTORIAL NOTES

---

presented by

Myra Spiliopoulou, Tanja Falkowski and Georgios Paliouras

September 17, 2007

Warsaw, Poland

Prepared and presented by:
Myra Spiliopoulou
Otto-von-Guericke University Magdeburg, Germany
Tanja Falkowski
Otto-von-Guericke University Magdeburg, Germany
Georgios Paliouras
National Center for Scientific Research "Demokritos", Greece

Discovering and Tracking User Communities

Myra Spiliopoulou[1], Tanja Falkowski[1], Georgios Paliouras[2]

[1] Otto-von-Guericke University Magdeburg, Germany

[2] National Center for Scientific Research "Demokritos"

The Presenters

Myra Spiliopoulou & Tanja Falkowski

- Work group KMD – Knowledge Management & Discovery
- Faculty of Computer Science
- Otto-von-Guericke-Universität Magdeburg
- Magdeburg, Germany
- http://omen.cs.uni-magdeburg.de/itikmd

Georgios Paliouras

- Software and Knowledge Engineering Lab.
- Institute of Informatics and Telecommunications
- National Center for Scientific Research "Demokritos"
- Athens, Greece
- http://www.iit.demokritos.gr/~paliourg

Presentation Outline

- Block 1: Community models
- Block 2: Three perspectives for community discovery
  - Similarity-based perspective
  - Interaction-based perspective
  - Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

Presentation Outline

- $\boxed{\text{Block 1: Community models}}$
- Block 2: Three perspectives for community discovery
  - Similarity-based perspective
  - Interaction-based perspective
  - Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

Notions of Communities

- Frequent informal definition of a *community*:

  Subset of vertices that has high density of edges within the group and a lower density of edges between groups
- A *Web community* is generally described as a substructure (subset of vertices) of a graph with dense linkage between the members of the community and sparse density outside the community [GibKleRag98]
- A community corresponds to a group of users who exhibit common behaviour in their interaction with the system [Orwant95]

Communities in Different Research Areas

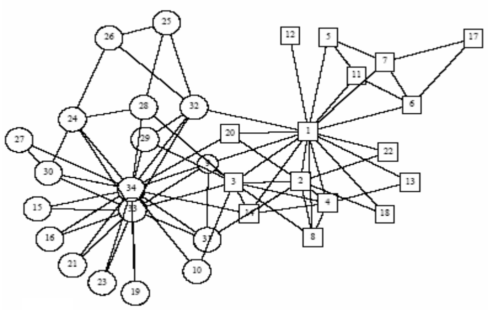
- Communities in Biology
  - Compartments in food webs
  - Functionally related genes
  - Functional groups in protein-protein interaction networks
- Communities in Social Sciences
  - (cohesive) subgroup of interacting individuals
- Communities in Computer Science
  - Set of Web Pages
  - Set of Servers
  - Group of Users

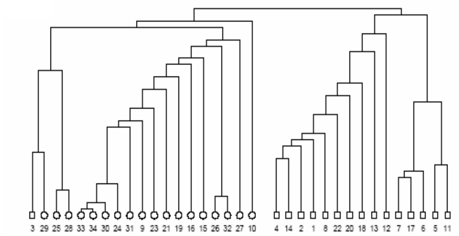
Communities in Friendship Networks

- Friendship network from Zachary Karate Club study
- Shown are two clusters:

  A: Actors associated with club administrator shown as circles

  B: Actors associated with instructor drawn as squares





Source: W.W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452–473 (1977)

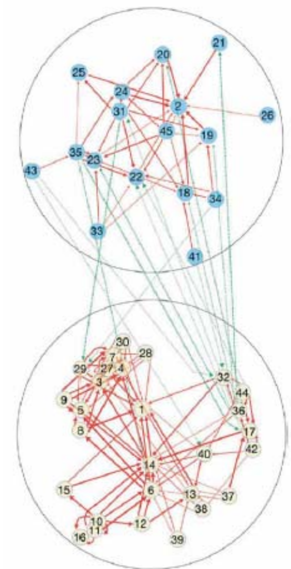
Compartments in Food Webs

- Predator-prey interactions (food web) in the Chesapeake Bay a large widely studied estuary in USA
- Shown are two compartments:

  A: pelagic taxa (species living in the water column)

  B: benthic taxa (species living at the bottom of a body of water; species living in sediments)

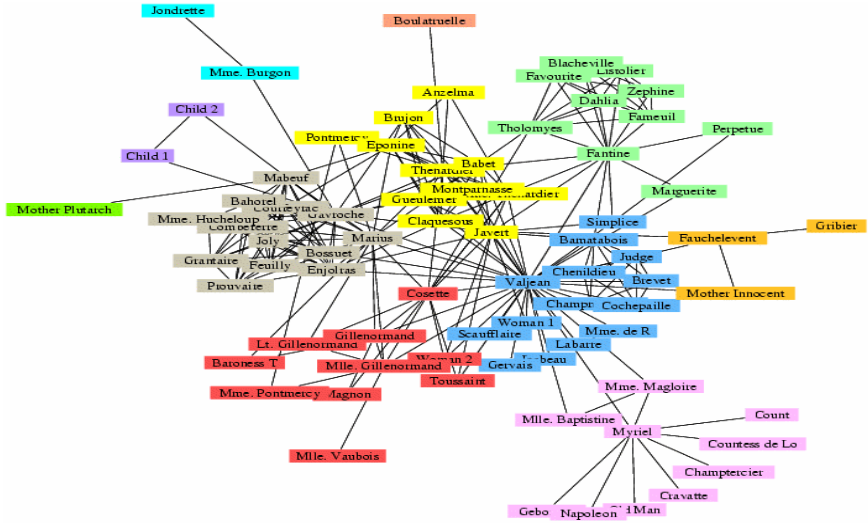
  - 65% of B's taxa interact with A; 30% of A's taxa interact with B
  - Placement of taxa indicates its role within the compartment



Source: S.R. Proulx, D.E.L. Promislow, P.C. Philipps, Network thinking in ecology and evolution, TRENDS in Ecology and Evolution, Vol. 20 No. 6, June 2005

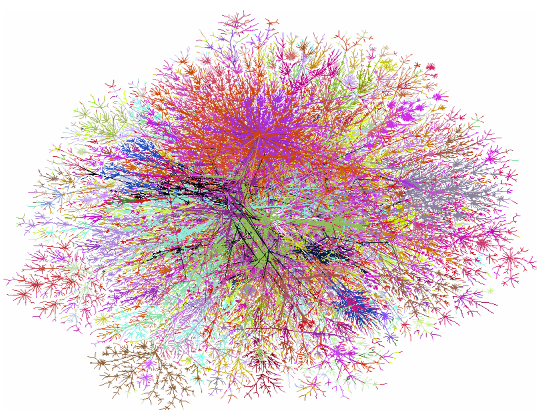
Communities in Co-appearance Network

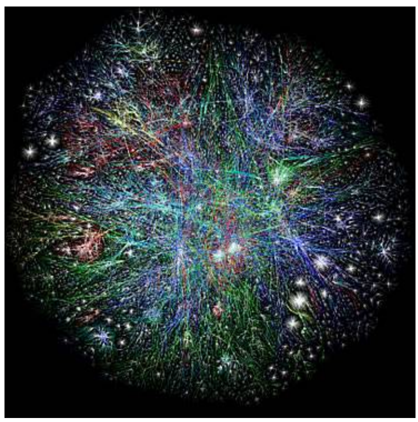
Les Miserables: Co-appearance in one or more scene



Source: M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004

Communities of Servers in the Internet





Source: http://www.cheswick.com/ches/map/gallery/wired.gif, April 23, 2007

Source: http://www.newscientist.com/article.ns?id=dn4434, April 23, 2007

References (Block 1)

- D. Gibson, J. M. Kleinberg, and P. Raghavan, Inferring Web Communities from Link Topology. In Proc. of ACM International Conference on Hypertext and Hypermedia (HT'98), 225-234, 1998
- M.E.J. Newman, M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113, 2004
- S.R. Proulx, D.E.L. Promislow, P.C. Philipps, Network thinking in ecology and evolution, TRENDS in Ecology and Evolution, Vol. 20 No. 6, June 2005
- W.W. Zachary, An information flow model for conflict and fission in small groups, Journal of Anthropological Research 33, 452–473, 1977
- J. Orwant: Heterogeneous Learning in the Doppelgänger User Modeling System. User Modelling and User Adapted Interaction, 2, 107-130, 1995

Presentation Outline

- Block 1: Community models
- Block 2: Three perspectives for community discovery
  - Similarity-based perspective
  - Interaction-based perspective
  - Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

Motivation

Similarity-Based User Communities

- User community: a group of similar people
  - Similar interests

    *Users(x,y,z) -> **like** (sports, stock market)*
  - Similar navigation behavior

    *Users(x,y,z) -> **visit**(sports news then football news)*

Similarity Based User Communities

- Early work: Site specific communities
  - Model common user interests.
  - Identify patterns in user navigation.
- Current work: Communities on the whole Web
  - Personalized Web directories (Yahoo!, ODP).
  - Include semantics in navigation patterns.

Site specific communities

- Stereotypes
- Communities of common interests
- Communities of common navigation

Site specific communities

- $\boxed{\text{Stereotypes}}$
- Communities of common interests
- Communities of common navigation

Stereotypes

- A stereotype is a means of describing the common characteristics of a class of users.
- It characterizes associates personal characteristics of the users with parameters of the system.

  Male users of age 20-30 are interested in sports and politics.
- Assumes registered users that provide personal/ demographic information,

  e.g. occupation, age, gender etc.

Stereotype construction

- Goal
  - Identify generic user models that associate stereotypical behavior with personal characteristics.
- Model
  - A stereotype corresponds to a class of users.
  - A set of attributes characterize the class.
- Approach:
  - Manual Construction.
  - Machine learning.

Stereotype construction (old fashion)

- Manual Construction
  - Predetermined stereotypes,
    e.g. child, adult, expert, etc.
  - The system collects personal information and assigns each user to a stereotype.
  - Stereotypes allows the system to anticipate some of the user's behavior and adapt its functionality.

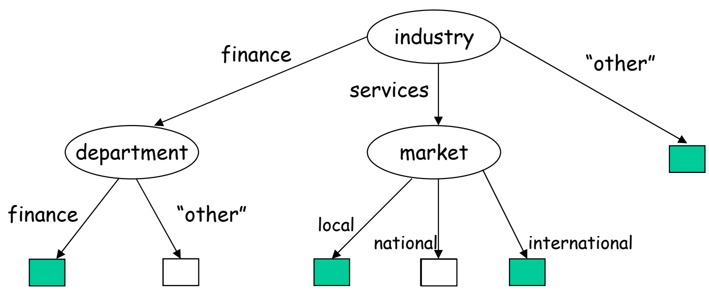
Stereotype construction (old fashion) – An Application: *Grundy Librarian System*: (Rich, CogSci79)

- The system suggests novels based on predetermined stereotypes.
- Each stereotype maintains statistics about the preferences of its users.
- Requires:
  - **Facets:**
    - Sets of user preferences, each associated with a value (or values). Stereotypes are simply collections of facet-value pairs that describe groups of system users.
  - **Triggers:**
    - Events (personal characteristics) that **activate** stereotypes.
  - **How?**
    - Ask questions and analyze answers.
    - Look for a trigger for a stereotype in the user's characteristics.

Stereotype discovery

- Machine learning
  - Associate behavioral patterns with personal information (supervised learning).
  - Algorithms:
    - Decision Trees (Paliouras et al, UM99)
      - Each decision tree is a stereotype modeling a system's variable, e.g. a category of news articles.
    - k-NN, naive Bayes, weighted feature vectors (Lock, AH06)
      - A stereotype corresponds to a set of features that represent each class.

Stereotype discovery - An example
Decision Trees (Paliouras et al, UM99)



IF (industry = finance AND department ≠ finance) OR (industry = services AND market = national)
THEN AND ONLY THEN the user is interested in company results

Stereotypes

- Applications:
  - News filtering and other IR tasks, digital libraries, electronic museums, etc.
- Problems
  - Hard to acquire accurate personal information.
  - Privacy issues.
- Solution: Restrict models to patterns in user behavior.
- We call these *user communities*.

Site specific communities

- ☑ Stereotypes
- Communities of common interests
- Communities of common navigation

Communities of common interests

- Goal
  - Identify similar users, i.e. users that share common interests.
- Model
  - Community models are clusters of users or clusters of common interests.
  - Each user belongs to one (or more if overlaps are allowed) communities.
- Approach
  - Collaborative Filtering.

Collaborative filtering

- Goal: Match a new user visiting a particular domain to a group of users in that domain with similar interests.
- Model:
  - A community is either a user-based or an item-based model of a group of users

    users(x,y,z) -> sports, stock market

    (business news, stock market) - > user(x), user(z)
- Algorithms:
  - memory-based learning,
  - model-based clustering,
  - item-based clustering.

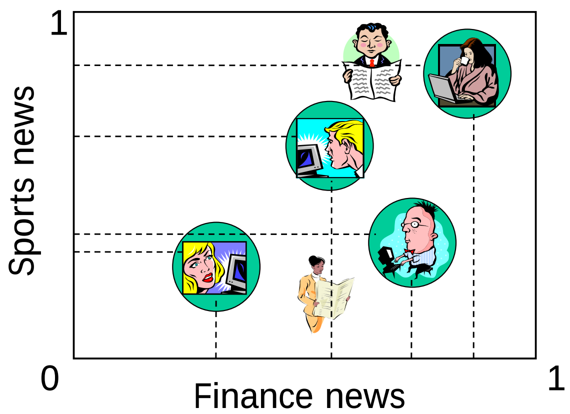
Memory-based learning

- Assumption
  - Exploit the whole corpus of users in order to construct a finite number of nearest neighbors close to the examined user.
- Algorithms
  - Mainly k-nearest neighborhood approaches.
- Model
  - The k-nearest neighbors correspond to an *ad-hoc* community.

Memory-based learning - (Herlocker et al, SIGIR99)

- Nearest-neighbor approach:
  - Construct a model for each user, based on the user's recorded preferences, e.g. item ratings.
  - Index the users in the space of system parameters, e.g. item ratings.
  - For each new user,
    - index the user in the same space, and
    - find the *k* closest neighbors.
    - create an ad-hoc community.
    - simple metrics to measure the similarity between users, e.g. Pearson correlation.
  - Recommend the items that the new user has not seen and are popular among the neighbors.

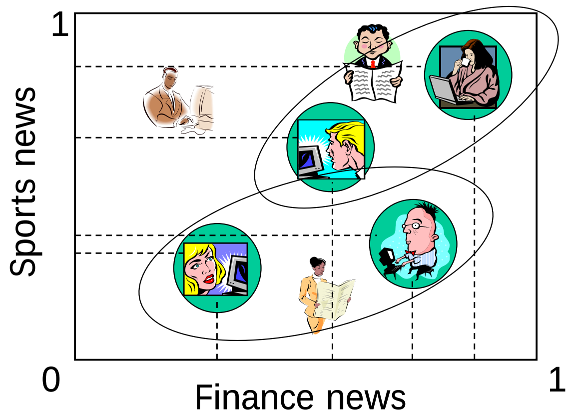
Memory-based learning



Model-based clustering

- Assumption
  - Machine learning techniques are applied, in order to create the user communities and then use the models to make predictions.
- Model
  - Community models: cluster descriptions.
  - Community models are global, rather than ad-hoc.

Model-based clustering



Model-based clustering

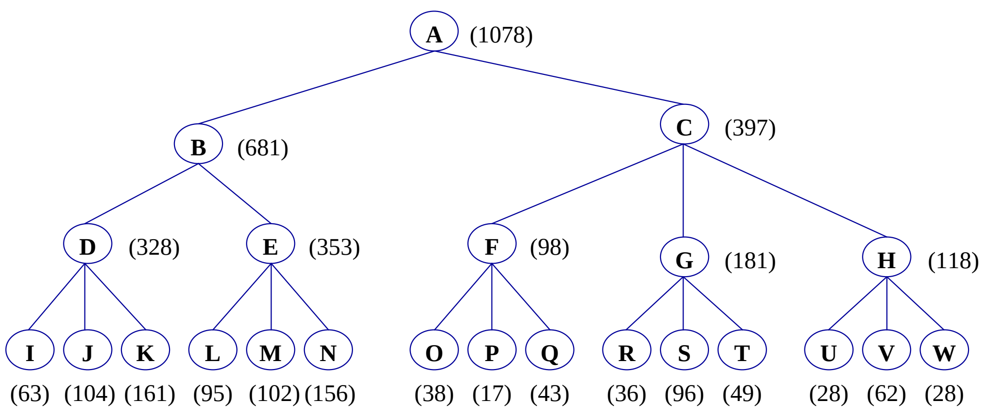
- Algorithms
  - K-Means and its variants.
  - Graph-Based clustering.
  - Conceptual clustering (COBWEB).
  - Statistical clustering (Autoclass).
  - Neural Networks (Self-Organizing Maps).
  - Model based clustering (EM-type).
  - BIRCH.
  - Fuzzy clustering.

Model-based clustering – Conceptual clustering (Paliouras et al, ICML00)

- Conceptual Clustering (COBWEB)
  - COBWEB generates a hierarchy of concepts.
  - Each concept is a cluster of objects.
  - Objects correspond to individual user models.
  - Concepts correspond to communities.
  - Similarity metric: *category utility*.
- Important: Each user in only one community.

Model-based clustering – Conceptual clustering (Paliouras et al, ICML00)

COBWEB Community hierarchy



Model-based clustering – Flexible Mixture Model (Si and Jin, ICML03)

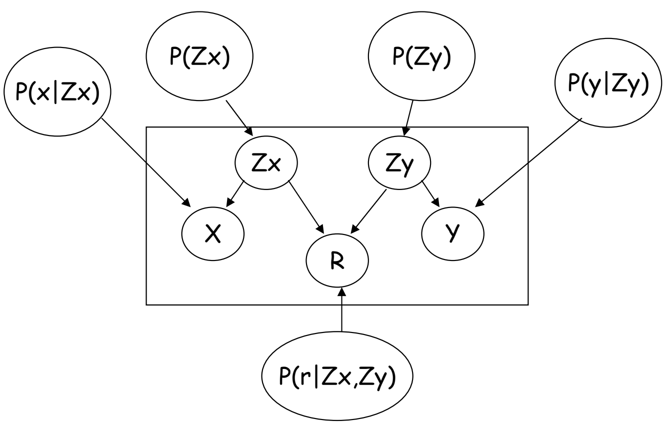
- Assume Z_X, Z_Y, latent variables indicating class membership for object (item) "x" and user "y" with multinomial distributions $P(Z_X)$, $P(Z_Y)$.
- The conditional probabilities: $P(X|Z_X)$, $P(Y|Z_Y)$, $P(r|Z_X, Z_Y)$ are the multinomial distributions for objects, users and ratings given Z_X, Z_Y.
- FMM model:

$$P(x,y,r) = \sum_{Zx,Zy} P(Zx)P(Zy)P(x \mid Zx)P(y \mid Zy)P(r \mid Zx, Z_Y)$$

- Expectation Maximization to calculate probabilities.
- Important: each user to more than one community.

Model-based clustering – Flexible Mixture Model (Si and Jin, ICML03)

Graphical Model Representation



Item-based clustering

- Goal
  - Identify behavior patterns in usage data, rather than user clusters.
- Model
  - Community models are clusters of items, e.g. Web pages.
  - Each item and each user belongs to one (or more if overlaps are allowed) communities.
- Algorithms
  - Similar to model-based clustering.

Item-based clustering - graph-based clustering (Paliouras et al, IwC02)

- Represent Web pages as bags of sessions:

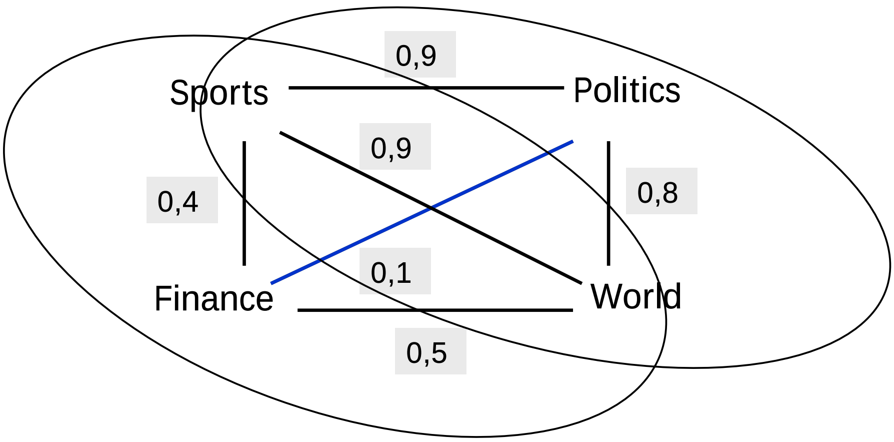
  [sports.html: ses1, ses12, ses123, ...]

  [racing.html: ses1, ses351, ...] ...
- Generate Graph G =< E, V,W_e,W_v >, where:

  V: pages, W_v freq. of occurrence,

  E: pairs of pages, W_e: freq. of co-occurrence.
- Remove edges according to a similarity threshold.
- Identify cliques in the graph.

Item-based clustering - graph-based clustering (Paliouras et al, IwC02)



Communities of Common Interests

- Applications
  - Query-based information retrieval.
  - Profile-based information filtering.
  - Adaptive Web sites.
  - Site reconstruction.
  - Recommendation.

Site specific communities

- ☑ Stereotypes
- ☑ Communities of common interests
- Communities of common navigation

Communities of common navigation

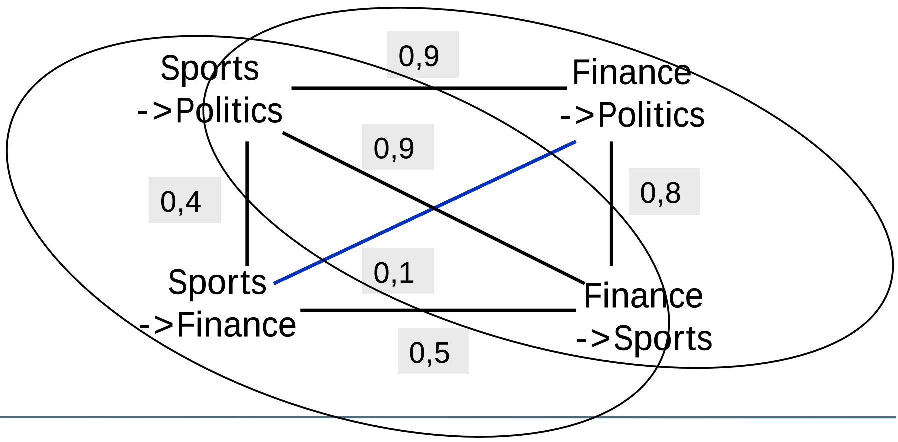
- Goal
  - Identify how users view the information.
  - Group users with similar navigation behavior.
- Model
  - Communities correspond to:
    - Sequential patterns, e.g. grammars.
- Algorithms
  - Sequential Pattern Discovery.
  - Grammatical Inference.

Communities of common navigation

- Sequential Pattern Discovery
  - Identifying navigational patterns, rather than “bag-of-page” models.
- Methods
  - Clustering transitions between pages.
  - First-order Markov models.
  - Probabilistic grammar induction.
  - Association-rule sequence mining.
  - Path traversal through graphs.
- Personal and community navigation models.

Communities of common navigation- Sequential Pattern Discovery (Paliouras et al, IwC02)

- Graph-based clustering; small modification of item-based clustering: an item is a transition between pages.



Communities of common navigation - Discovering Grammatical Models (Karambatziakis et al, ICGI04)

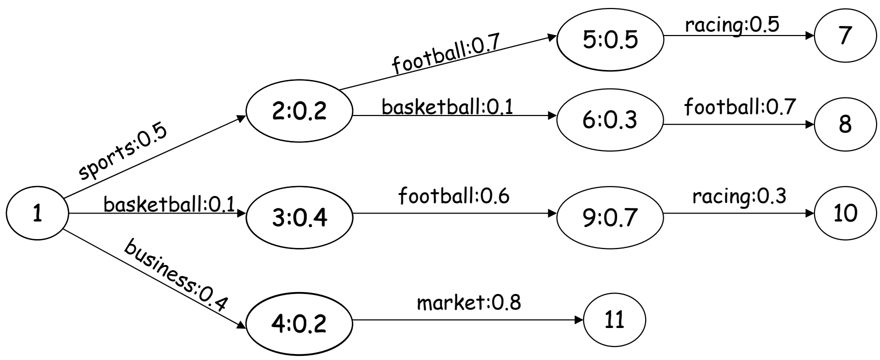
- Each Web page is a terminal symbol of a language L.
- Each user session is a string of the language.
- Assume strings are generated by an unknown grammar, modeled by a deterministic probabilistic Stochastic Finite Automaton (SFA).
- Use grammatical inference to discover the automaton.

Communities of common navigation - Discovering Grammatical Models (Karambatziakis et al, ICGI04)

- Discovering Grammatical Models
  - Represent the data as a tree, in particular a PPTA: probabilistic prefix tree automaton.
  - Iteratively merge compatible states, preserving determinism.
  - Compatibility = similar outward transitions.
  - Heuristic search of the space of compatible states.

Communities of common navigation - Discovering Grammatical Models (Karambatziakis et al, ICGI04)

A simple example



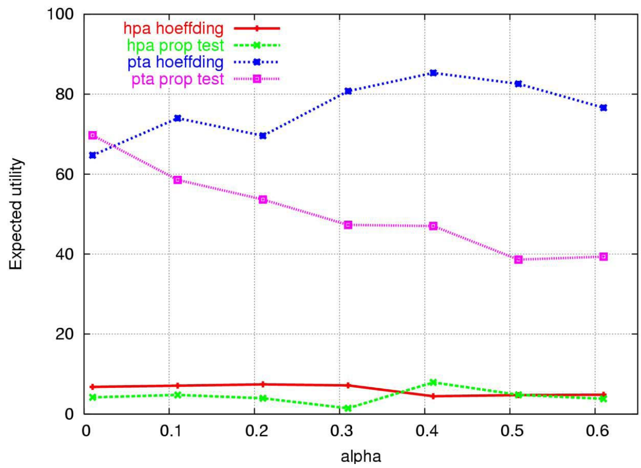
Communities of common navigation

- Discovering grammatical models – Experiments:
  - Recommendation on two large Web sites:
    MSWeb and a portal on chemistry.
  - Evaluation process:
    1. Build model on part of the usage data.
    2. Hide the last page in each test session.
    3. Trace observed path on the automaton.
    4. Build recommendation list from current node's children.
  - Evaluation measure (expected utility):

$$EU_a = \sum_{j=0}^{n-1} \frac{v_{aj}}{2^{j/h}}$$

Communities of common navigation

- Results



Communities on the whole Web

- Motivation: The challenge of acquiring user models on the Web.
  - Usage data is voluminous.
  - Web structure is unknown and complex.
  - The users' interests, knowledge and behavior is diverse.
  - The thematic coverage of the data is very broad.

Communities on the whole Web

- Model similar interests of Web users:
  - Community Web directories (Yahoo!, ODP).
- Model similar navigation behavior on the Web:
  - Content-aware navigation user modeling with GI.

Communities on the whole Web

- Community Web Directories
- Web Navigation Models

Communities on the whole Web

- Community Web Directories
- Web Navigation Models

Community Web directories

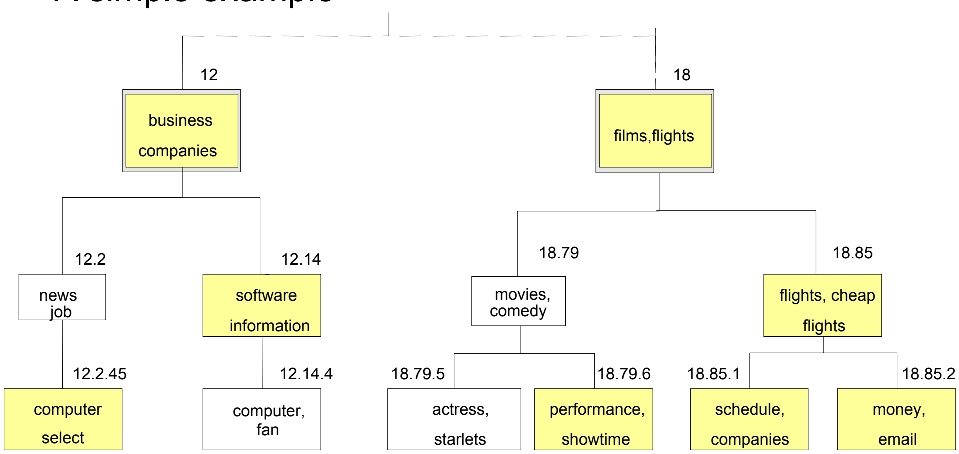
- Personalization of and with Web Directories
- Model:
  - Analyzing usage data collected by the proxy servers of an Internet Service Provider (ISP).
  - Construction of user community models.
  - Construction of usable Web directories that correspond to the interests of user communities.
- Algorithms:
  - Graph-Based Clustering.
  - Probabilistic Latent Semantic Analysis (PLSA).

Community Web directories

- Off-line user modeling:
  - Map user sessions on the directory categories, i.e. each session becomes a small subdirectory.
  - Create community Web directories.
  - Prune non-representative branches.
  - Remove redundant nodes, e.g. those without siblings.
- On-line use of community directories
  - Personal Web directories constructed by assigning users to community directories and merging them.
  - Personalized directories are small and provide quick access to interesting information.

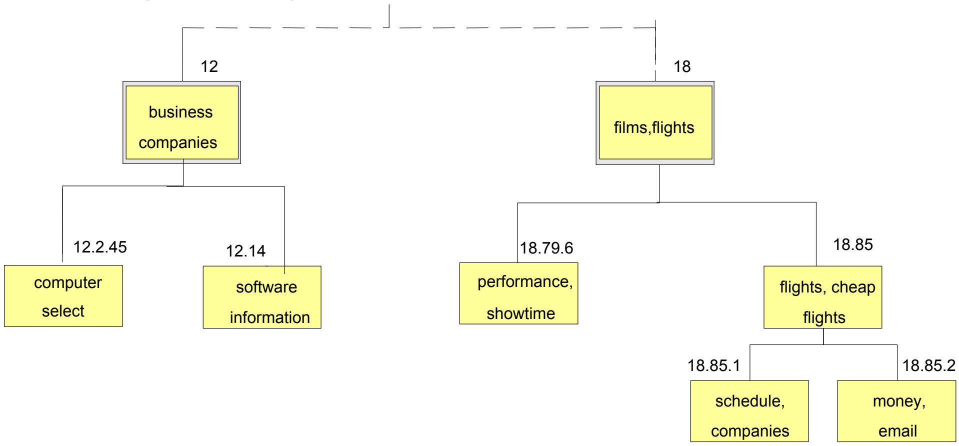
Community Web directories

- A simple example



Community Web directories

- A simple example



Community Web directories – graph-based clustering (Pierrakos et al., EWMF03)

- A modified version of the method used for Web sites:
  - Each directory category k_i becomes a node in the graph.
  - Each page p_j is assigned a set K_j of categories, including all ancestors.
  - For each occurrence of page p_j increase the weight of all $k_{ji} \in K_j$.
  - For each co-occurrence of p_j and p_l increase the weight of all (k_{ji}, k_{lm}), $k_{ji} \in K_j$, $k_{lm} \in K_l$ edges.
  - Reduce connectivity of the graph and find cliques.
  - Construct a community directory for each clique.

Community Web directories - latent-factor modeling (Pierrakos et al., UM05)

- Assume: a session u_i is due to a latent factor z_k, characterizing a community.
- Model the probability $P(u_i, c_j)$, where c_j a directory category:

$$P(u_i, c_j) = \sum_k P(z_k) P(u_i | z_k) P(c_j | z_k)$$

- Use Expectation Maximization to estimate the probabilities from the data.
- Construct a community directory for each factor, using the most representative categories: $P(c_j|z_k) > T_z$.

Community Web directories

- Evaluation
  - 781,069 records from ISP proxy server log.
  - After cleaning and sessionization: 2,253 sessions
  - Initial Web directory constructed with agglomerative document clustering (998 nodes).
  - Repeated split of the data for modeling and evaluation.
  - Hide last page from each evaluation session.
  - Use observed pages to construct personal directory.

Community Web directories

- Evaluation Metrics:
  - Coverage: percentage of hidden pages covered by the personalized directories.
  - User Gain:
    - Position hidden page p_i in the directory.
    - Measure *click path*:

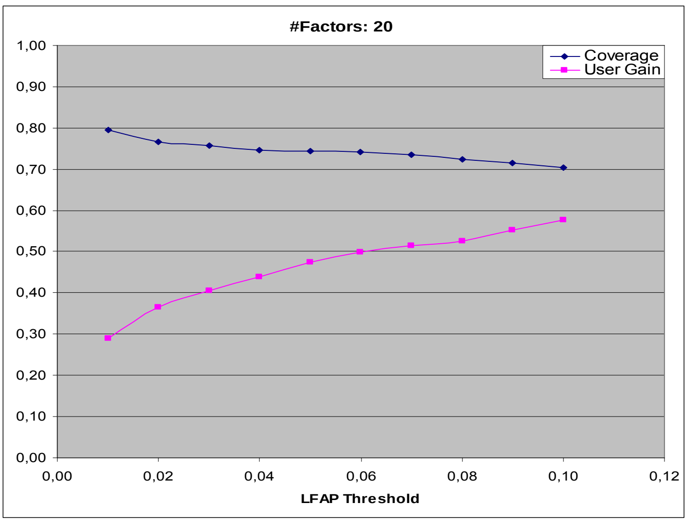
$$CP_i = \sum_{j}^{depth} j \times branch\_factor_j$$

    - Measure average gain over original directory:

$$UG = \sum_i \frac{CP_i^{gen} - CP_i^{pers}}{CP_i^{gen}}$$

Community Web directories

- Results



Communities on the whole Web

- ☑ Community Web Directories
- Web Navigation Models

Modeling navigation on the Web

- Model how people navigate the Web.
- Acquire models from Web usage data, e.g. ISP.
- Can we apply the same methods as for a Web site?
- Statistics of Web page co-occurrence do not allow that.
- Approach: model also content-based page similarity.

Modeling navigation on the Web – Content-Aware Navigation User Modeling (Korfiatis et al. AAI08)

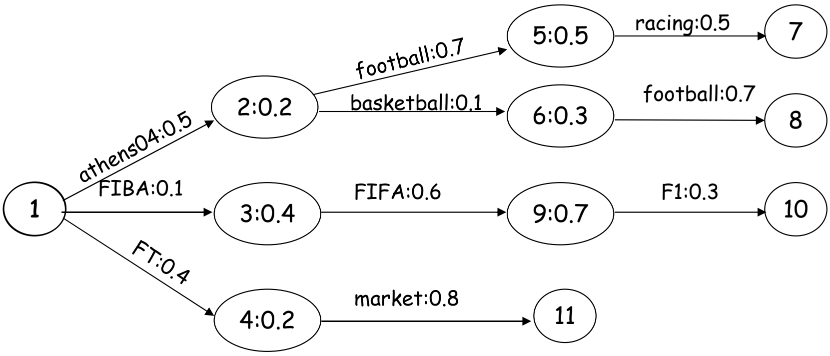
- Stick to grammars as navigation models.
- Key: each state is a cluster of the pages that lead to it.
- Each page (and page cluster) is represented as a word-frequency vector: [goal=0.2,shot=0.1,basket=0,money=0.05].
- We can measure state compatibility by combining transition probabilities with vector similarity, e.g. using the cosine metric.

Modeling navigation on the Web

- Content-Aware Navigation User Modeling with GI
  - Extend state compatibility to use content similarity:
  - Measure usage and content similarity:
    $u(s_1, s_2)$, $c(s_1, s_2)$.
  - Reject merge if $u(s_1, s_2) < T_u$ or $c(s_1, s_2) < T_s$.
  - Normalize thresholds using the metric distributions in the PPTA.
  - Combine by min, max, or weighted average.
  - Search for most compatible pair of states as usual.

Modeling navigation on the Web

- A simple example



Modeling navigation on the Web

- On line recommendation process
  - Modify recommendation process to use content similarity:
  - Given a state s_i, with children S_i, and the next observed page of the user's session *a*, select $argmax_j$ $sim(a,s_{ij})$.
  - If $argmax_j$ $sim(a, s_{ij})$ < Tsim return to start state.
  - At the end of the observed path, build recommendation list combining:
    - The transition probability to the final state's children.
    - The distance of each page in a state to the state's centroid.

Modeling navigation on the Web

- Evaluation:
  - Data: the ISP data used for personalized directories.
  - Modification of the Expected Utility measure:

$$EU_a = \sum_{j=0}^{n-1} \frac{sim(a, p_j)}{2^{j/h}}$$

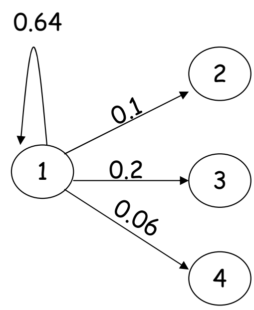
  - Comparison to content-only recommendation:
    - Store all pages in the modeling phase.
    - Score stored pages, according to average content distance from the observed path.
    - Produce a list of the n top-scoring pages.

Modeling Navigation on the Web

- Results:

Method	EU
CANUMGI-A	8.57
CANUMGI-B	21.72
CANUMGI-C	20.59
CONTENT	24.25

Does the navigation model help?



Navigation Sequences are thematic

References Block 2 – Similarity-based perspective

- G. Paliouras, V. Karkaletsis, C. Papatheodorou and C.D. Spyropoulos, "Exploiting Learning Techniques for the Acquisition of User Stereotypes and Communities," Proceedings of the International Conference on User Modeling (UM), CISM Courses and Lectures, n. 407, pp. 169-178, Springer-Verlag, 1999.
- Lock, Z. and Kudenko, D., "Interaction Between Stereotypes", In Proc. of International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems (AH2006), 2006.
- Herlocker, J., Konstan, J., Borchers, A., and Riedl, J, "An Algorithmic Framework for Performing Collaborative Filtering". In Proc. 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA, USA 230-237, 1999.
- G. Paliouras, C. Papatheodorou, V. Karkaletsis and C.D. Spyropoulos, "Clustering the Users of Large Web Sites into Communities," Proceedings of the International Conference on Machine Learning (ICML), pp. 719-726, Stanford, California, 2000.
- L. Si and R. Jin, A Flexible Mixture Model for Collaborative Filtering, In the Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)

References Block 2 – Similarity-based perspective

- G. Paliouras, C. Papatheodorou, V. Karkaletsis and C.D. Spyropoulos, "Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques,". Interacting with Computers, v. 14, n. 6, pp. 761-791, 2002
- N. Karampatziakis, G. Paliouras, D. Pierrakos, P. Stamatopoulos, "Navigation pattern discovery using grammatical inference," In Proceedings of the 7th International Colloquium on Grammatical Inference (ICGI), Lecture Notes in Artificial Intelligence, n. 3264, pp. 187 - 198, Springer, 2004
- D. Pierrakos, G. Paliouras, C. Papatheodorou, V. Karkaletsis, M. Dikaiakos, "Web Community Directories: A New Approach to Web Personalization," In Berendt et al. (Eds.), "Web Mining: From Web to Semantic Web", Lecture Notes in Computer Science, n. 3209, pp. 113 - 129, Springer, 2004
- D. Pierrakos, G. Paliouras, "Exploiting Probabilistic Latent Information for the Construction of Community Web Directories," In Proceedings of the International User Modelling Conference (UM), Edinburgh, UK, July, Lecture Notes in Artificial Intelligence, n. 3538, pp. 89-98, Springer, 2005
- Korfiatis, G and Paliouras, G. "Modeling Web Navigation using Grammatical Inference", to appear in AAI

Presentation Outline

- Block 1: Community models
- Block 2: Three perspectives for community discovery
  - ☑ Similarity-based perspective
  - Interaction-based perspective
  - Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

Block 2: Interaction-based Community Detection

- Types of Interaction
  - Communication
    - face-to-face
    - telephone
    - email
    - …
  - Recommendation
  - Co-Authoring
  - …

Graph-Representation of Interaction Networks

- Possible representation of networks are graphs
- Graph $G=(V,E)$ with vertices (nodes) V and edges (links) E
- Studying global characteristics of graphs (using statistical measures)
- Studying the topology of graphs, such as subgroups (subset of connected nodes)

Cohesive Subgroups in Social Sciences

- Definition based on relative strength, frequency density or closeness of ties within the subgroup and
- relative weakness, infrequency, sparseness, or distance of ties from subgroup members to others

1. Methods based on properties of ties within the subgroup
2. Methods based on comparison of ties within the subgroups to ties outside the group

Cohesive Subgroups
in non-directed networks

A **cohesive subgroup** is a subset of actors among whom there are relatively strong, direct, intense or frequent ties

- Subgroups based on complete mutuality: *Cliques*
  - Maximal complete subgraph of three or more nodes (i.e. all nodes are adjacent to each other)
- Subgroups based on reachability and diameter: *n-cliques*
  - Maximal subgraph in which the largest geodesic distance between any two nodes is no greater than *n*
- Subgroups based on nodal degree: *k-plexes*, *k-cores*
  - A *k-plex* is maximal subgraph containing *s* nodes in which each node is adjacent to no fewer than *s-k* nodes in the subgraph
  - A *k-core* is a subgraph in which each node is adjacent to at least *k* other nodes in the subgraph

Community Detection Methods and Applications
Based on Graphs of Interactions

- Maximum flow minimum cut
- Hierarchical divisive clustering
- Hyperlink-Induced Topic Search (HITS) and PageRank

Maximum-flow minimum cut theory
Algorithm: Idea

- Given a directed graph $G=(V,E)$, with edge capacities $c(u,v) \in Z^+$, and two vertices $s, t \in V$.
- Find the *maximum flow* that can be routed from the source s to the sink t that obeys all capacity constraints.
- A *minimum cut* of a network is a cut whose capacity is minimum over all cuts of the network
- *Max-flow-min-cut theorem* of Ford and Fulkerson (1956) proves that maximum flow of the network is identical to minimum cut that separates s and t.

Maximum-flow minimum cut theory:
Algorithm: Ford-Fulkerson Method

- Method to solve the maximum-flow problem
- **Residual Capacity:** Additional net flow we can push from u to v before exceeding the capacity $c(u,v)$

  $c_f(u,v) = c(u,v) – f(u,v)$
- **Augmenting path**: Path from source s to sink t along which we can push more flow
- Repeatedly augmenting the flow until the maximum flow has been found
- A **cut** (S,T) of the flow network G is a partition of V into S and $T = V\text{-}S$ such that $s \in S$ and $t \in T$

Maximum-flow minimum cut theory:
Algorithm: Ford-Fulkerson Method

```
Ford-Fulkerson(G,s,t)
  1 for each edge (u,v) ∈ E[G]
  2    do f[u,v] ← 0
  3       f[v,u] ← 0
  4 while there exists an augmenting path p from s
       to t in the residual network G_f
  5    do c_f(p) ← min{c_f(u,v): (u,v) is in p}
  6       for each edge (u,v) in p
  7          do f[u,v] ← f[u,v] + c_f(p)
  8             f[v,u] ← -f[u,v]
```

- Lines 1-3 initialize the flow
- While loop of lines 4-8 repeatedly finds augmenting path *p* in G_f and augments flow *f* along *p* by the residual capacity $c_f(p)$
- When no augmenting paths exits, the flow is maximum flow

Application „Identification of Web Communities“
[Flake, Lawrence & Giles, 2000]

- Definition of Community: A Web community is a collection of Web pages in which each member page has more hyperlinks within the community than outside the community.
- Goal: Finding topologically related Web sites (e.g. to reduce the number of Web sites to index)
- Model: Two Web sites are connected via a directed edge if one site links to the other
- Algorithm: Focused-crawl based on max-flow analysis

Application „Identification of Web Communities“: Algorithm
[Flake, Lawrence & Giles, 2000]

```
FOCUSED-CRAWL(G,s,t)
while # of iterations is less than desired do
  Perform maximum flow analysis of G, yielding community C.
  Identify non-seed vertex, v*∈C, with the highest in-
   degree relative to G.
  for all v ∈ C with in-degree equal to v*,
         Add v to seed set
         Add edge (s, v) to E with infinite capacity
  end for
     Identify non-seed vertex, u*, with the highest out-
   degree relative to G
  for all u ∈ C with out-degree equal to u*,
          Add u to seed set
          Add edge (s, u) to E with infinite capacity
  end for
     Re-crawl so that G uses all seeds
     Let G reflect new information from the crawl
end while
```


Application „Identification of Web Communities“: Results
[Flake, Lawrence & Giles, 2000]

- The authors test their algorithm with three different groups of initial Web pages. Each retrieved community is closely related to the interested field:
  - Support Vector Machine Community
    - Graph Size: 11,000
    - Community Size: 252
    - Results: strongly related to SVM research
  - The Internet Archive Community
    - Graph Size: 7,000
    - Community Size: 289
    - Results: closely related to the mission of the Internet Archive
  - The “Ronald Rivest” Community
    - Graph Size: 38,000
    - Community Size: 150
    - Results: closely related to Ronald Rivest’s research

Community Detection Methods and Applications

- ☑ Maximum flow minimum cut
- Hierarchical divisive clustering
- Hyperlink-Induced Topic Search (HITS) and PageRank

Hierarchical Divisive Clustering

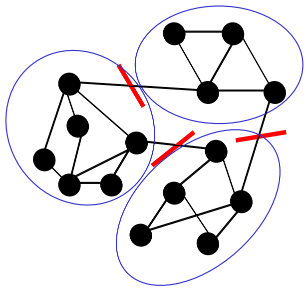
Core idea:

- The network is partitioned into groups with hierarchical divisive clustering
- The partitioning is done by removing edges according to the edge betweenness criterion of (Girvan & Newman, 2002)
- The output of the clustering algorithm is a dendrogram
- The dendrogram is "cut" at some level. The clusters are the graph partitions at this level
- The cut is performed according to a quality measure of (Newman & Girvan, 2004)

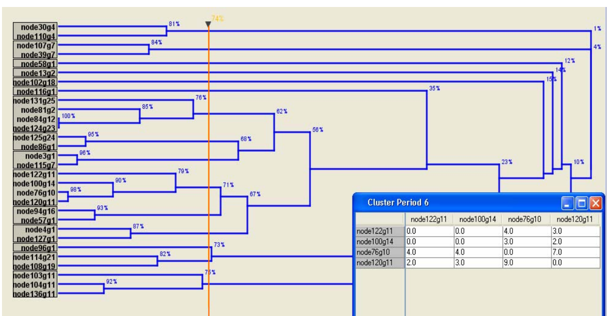
Hierarchical Divisive Clustering
Algorithm

- When a graph is made of tightly bound clusters, loosely interconnected, all shortest paths between clusters have to go through the few inter-cluster connections
- Inter-cluster edges have a high *edge betweenness*
- The *edge betweenness* of an edge e in a graph *G*(*V*,*E*) is defined as the number of shortest paths between all pairs of nodes along it

```
EDGE BETWEENNESS CLUSTERING (G)
    repeat until no more edges in G
        Compute edge betweenness for
        all edges
        Remove edge with highest
        betweenness
    end
```

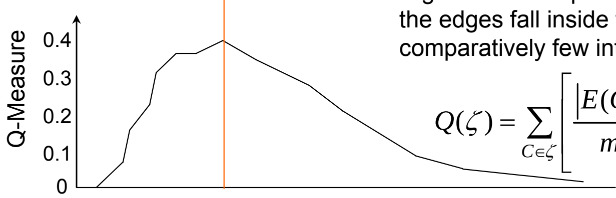


Hierarchical Divisive Clustering
Quality Measure



The dendrogram

Quality-Measure [Newman & Girvan, 2004].
A good network partition is obtained if most of the edges fall inside the communities, with comparatively few inter-community edges.



$$Q(\zeta) = \sum_{C \in \zeta} \left[\frac{|E(C)|}{m} - \left(\frac{\sum_{v \in C} \deg(v)}{2m} \right)^2 \right]$$

Application „Community Structures from Email“
[Tyler, Wilkinson, Huberman, 2003]

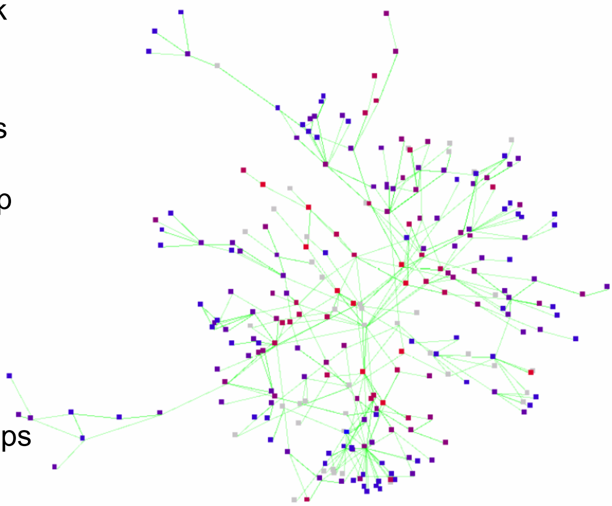
- Goal: Finding groups of people (communities of practice) interacting via email; draw inferences about the leadership of an organization from its communication data
- Model: Nodes represent users; two users are connected via a directed edge if they exchanged at least 30 emails and each user had sent at least 5 emails to the other
- Algorithm: Hierarchical divisive edge betweenness clustering with modifications

Application „Community Structures from Email“: Data Set
[Tyler, Wilkinson, Huberman, 2003]

- 185,773 emails between 485 HP Labs employees (November 2002 – February 2003)
- Emails to or from external destinations are removed
- Messages sent to a list of more than 10 recipients have been removed (such as lab-wide announcements)
- Graph consisted of 367 nodes connected by 1110 edges
- 66 communities were detected; largest consisted of 57 individuals; mean community size 8.4; $\sigma = 5.3$
- 49 of 66 communities consisted of individuals entirely within one lab or unit

Application „Community Structures from Email“: Results
[Tyler, Wilkinson, Huberman, 2003]

- Structure of email network bears resemblance with structure of organization
- Graph visualization shows that organizational leadership tends to end up in the center of the graph (red dots)
- Results were validated in interviews
- Communities reflect departments, project groups or discussion groups



Community Detection Methods and Applications

- ☑ Maximum flow minimum cut
- ☑ Hierarchical divisive clustering
- Hyperlink-Induced Topic Search (HITS) and PageRank

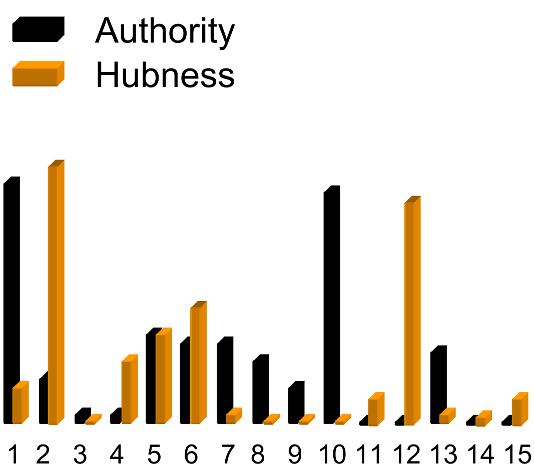
HITS Algorithm
[Kleinberg, 1999]

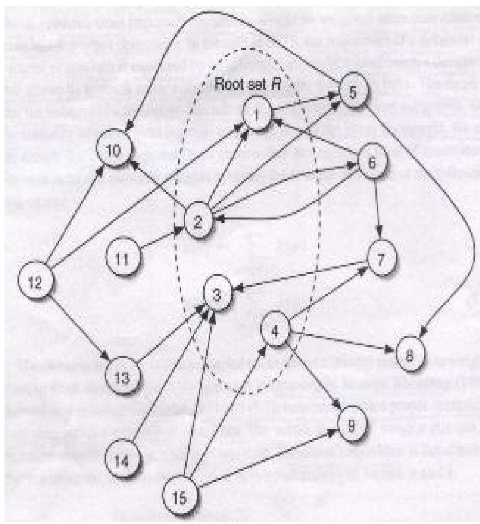
- Idea: Authorities are pages that are linked by many hubs. Hubs are pages that link to many authorities. HITS retrieves the bipartite core of a subgraph.
- Model: Collection *V* of hyperlinked pages as a directed graph *G* = (*V*, *E*): the nodes correspond to the pages, and a directed edge (*p*, *q*) indicates the presence of link from *p* to *q*. The authority score *a* and hub score *h* for a page *p* is calculated as follows

$$a_p = \sum_{q:(q,p)\in E} h_q \qquad h_p = \sum_{q:(p,q)\in E} a_q$$

- Goal: Detecting clusters of (topically) related pages

HITS Algorithm: Example





Source: Pierre Baldi, Paolo Frasconi, Padhraic Smyth, *Modeling the Internet and the Web*, Wiley, 2003

Page Rank
[Brin, Page, 1998]

- Idea:
  - Link analysis algorithm assigns numerical weight to each element of a hyperlinked set of documents such as the WWW
  - Assumptions: Link to page reflects “quality” and important pages link most likely to other important pages
- Model:
  - Collection *V* of hyperlinked pages as a directed graph *G* = (*V*, *E*): the nodes correspond to the pages, and a directed edge (*p*, *q*) indicates the presence of link from *p* to *q*.
- Goal:
  - Measure the relative importance of a page within the set
  - Importance of page affects other pages and depends on the importance of them → recursively

Calculation of Page Rank
[Brin, Page, 1998]

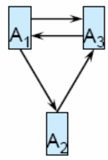
- The PageRank-value PR_i of page *i* is obtained from the weights of all pages that link to *i*. The PageRank of page *j* is divided among all the C_j outbound links. Thus, the PageRank of page *i* is calculated as follows:

$$PR_i = d \sum_{\forall j \in \{(j,i)\}} \frac{PR_j}{C_j} + (1-d)$$

- *d*=[0,1] is the dampening factor that is subtracted from the weight (1-*d*) of each page and distributed equally to all pages. It is generally assumed that the damping factor will be set around 0.85.

Page Rank: Example

1. Initialize *PR*; *d*=0,5
2. Value for *n* results from value of *n-1* using the *PageRank* equation
3. Repeat the calculation until values converge

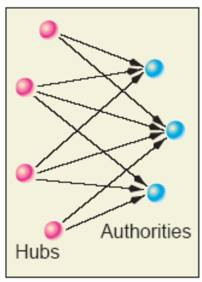


	A_1	A_2	A_3
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.7656	1.1484
3	1.0742	0.7686	1.1528
4	1.0764	0.7691	1.1537
5	1.0768	0.7692	1.1538
6	1.0769	0.7692	1.1538
7	1.0769	0.7692	1.1538
8	1.0769	0.7692	1.1538
9	1.0769	0.7692	1.1538

$$PR_i = d \sum_{\forall j \in \{(j,i)\}} \frac{PR_j}{C_j} + (1-d)$$

HITS and PageRank: Detecting Communities

- PageRank and HITS relate to spectral graph partitioning
- Characteristic patterns of hubs and authorities can be used to identify communities of pages on the same topic (see Figure right)
- Several modifications of HITS algorithm are proposed to detect communities in the Web
  - Gibson, D., Kleinberg, J., M., Raghavan, P., Inferring Web Communities from Link Topology, In Proc. of the 9th ACM Conference on Hypertext and Hypermedia, 225-234, 1998
  - Kumar, R., Raghavan, P., Rajagopalan, S., Trawling the Web for emerging cybercommunities, Computer Networks, Vol. 31, No. 11-16, 1481-1493, 1999



Community Detection Methods and Applications

- Maximum flow minimum cut
- Hierarchical divisive clustering
- Hyperlink-Induced Topic Search (HITS) and PageRank

References (Block 2 Part 2)

- Brin, S. and Page L., The anatomy of a large-scale hypertextual Web search engine". In. Proc. of 7th Interntl. Conference on World Wide Web, 107-117, 1998
- Flake, G.W., Lawrence, S., and Giles, C.L., Efficient Identification of Web Communities, In Proc. of Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000
- Ford Jr., L.R. and Fulkerson, D.R., Maximal flow through a network. *Canadian J. Math*., 8:399–404, 1956
- Girvan, M. and Newman, M.E., Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA*, **99**, 7821-7826, 2002
- Kleinberg, J. Authoritative Sources in a Hyperlinked Environment, Journal of the ACM, 46, 5, 604 –632, 1999
- Kleinberg, J. and Lawrence, S., The Structure of the Web, SCIENCE VOL 294, 1849-50, 2001
- Leskovec, J., Adamic, L.A., Huberman, B.A., *The Dynamics of Viral Marketing*, ACM Transactions on the Web, 1, 1, 2007
- Newman, M. and Girvan, M., Finding and evaluating community structure in networks, *Physical Review E* 69(026113), 2004
- Tyler, J.R., Wilkinson, D.M. and Huberman, B.A., Email as spectroscopy: automated discovery of community structure within organizations, Kluwer, 81-96, 2003

Presentation Outline

- Block 1: Community models
- Block 2: Three perspectives for community discovery
  - ☑ Similarity-based perspective
  - ☑ Interaction-based perspective
  - ☐ Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

An Impact-Oriented View upon Communities

- Tracing the influential members in a group of individuals
- Patterns of influence in a social network
- Being influenced to join a community

Influential individuals in marketing applications
Assessing *network value* in (Domingos & Richardson, KDD'01)

- In *direct marketing* applications, a marketing action towards a customer is performed
  if the cost of the action is lower than the expected profit.
- The expected profit is traditionally computed upon the *intrinsic value* of the customer – the profit from purchases of this customer.
- Domingos & Richardson proposed to consider also the *network value* of a customer – the profit from purchases done by other people, as the result of the *influence* of this customer.
  - Since then, much attention has been drawn to the influential members of social networks (markets or not).

Viral Marketing

The method of (Domingos & Richardson, KDD'01)
Modeling a market as a social network

- Actions of relevance for a customer X:
  - be the target of a marketing action
  - buy a product
- A customer X has neighbours:
  - A *neighbour* of X is a customer that directly influences X.
  - A customer X' influences X with some likelihood, which
  - depends on the marketing action directed to X'
    and on the attributes of the product.
- We compute the probability that X buys a product, given
  - the attributes of the product *Y* and
  - the marketing actions *M* directed to the neighbours of X
  - and the spreading nature of influence.

$$P\left(X \mid N(X), Y, M\right)$$

The method of (Domingos & Richardson, KDD'01)
Customer network value in a market

- The *Intrinsic Value* of a customer corresponds to the *expected lift in profit* achieved by directing a marketing action to this customer and ignoring the customer's influence upon others.
- The *global lift in profit* for a selection S of customers corresponds to their intrinsic values PLUS the expected lift in profit effected through their influence upon others.
  - The *Total Value* of a customer is the difference between the global lift in profit when including vs excluding this customer from S.
  - The *Network Value* of a customer is the difference between her *Total Value* and *Intrinsic Value*.

The method of (Domingos & Richardson, KDD'01)
The viral marketing problem in a social network

- Objective is to find the selection S of customers that maximizes the global lift in profit.
- The problem is intractable.
- Possible heuristics:
  - Consider each customer / marketing action only once.
  - Consider a customer for a marketing action only if this improves the previous value of the global lift in profit.
  - Launch a hill-climbing method.
- Experiments on EasyMovie (simulating a market):
  - The mass-marketing strategy yielded negative profit.
  - Direct marketing with the second heuristic turned to perform comparably to the hill-climbing method.

The authors consider the equivalent objective of determining the optimal set of direct marketing actions.

Influence of
the method of (Domingos & Richardson, KDD'01)

The topic "influence of individuals in viral marketing" enjoyed (has triggered ?) much further work, including

- More general models for viral marketing with Markov random fields by (Domingos et al)
- Cascades of influence for viral marketing and for social networks in general by (Kleinberg et al)
  - Modeling spread of influence (KDD'03)
  - Cascades in a recommendation network (PAKDD'06)
  - Cascades and group evolution in research networks (KDD'06)
  - ...

Spread of influence in a network
Problem formalization and analysis in (Kempe et al, KDD'03)

We observe a Social Network as a medium for the spread of an idea, innovation, item I:

- Understand the network diffusion processes for the adoption of the new I.

  Well-studied problem in social sciences, among else for the acceptance of medical innovations

- Given is a network N.
  We want to promote a new I to that set S of individuals, such that a maximal set of further adoptions will follow.

  "Influence Maximization Problem"
  New formal problem ***p*** posed by Domingos and Richardson

Spread of influence in a network
Problem formalization and analysis in (Kempe et al, KDD'03)

We observe a Social Network as a medium for the spread of an idea, innovation, item I:

- Understand the network diffusion processes for the adoption of the new I.

  Well-studied problem in social sciences, among else for the acceptance of medical innovations

- Given is a network N.
  We want to promote a new I to that set S of individuals, such that a maximal set of further adoptions will follow.

  "Influence Maximization Problem"
  New formal problem *p* posed by Domingos and Richardson

Basic Network Diffusion Models
(source: Kempe et al, KDD'03)

The social network is modeled as directed graph G
a node of which can be

- active := adopter of the new I
- inactive

Assumption, to be lifted later

The *progress* of activation is observed, in which an inactive node can become active but not vice versa.

The tendency of a node to become active increases monotonically with the number of its active neighbours.

Two basic models for this progress:

- Linear Threshold Model
- Independent Cascade Model

Basic Network Diffusion Models
(source: Kempe et al, KDD'03)

Linear Threshold Model:

- A node v is associated with an activation threshold τ_v.
- An active neighbour w of v influences v by a value $b_{w,v}$.
- The diffusion process unfolds in discrete steps.
- At iteration j, node v becomes active if and only if the received influence from its active neighbours exceeds the own threshold. $\sum_{w \in activeNeighbours(v,j)} b_{w,v} \geq \tau_v$

The activation threshold reflects the latent tendency of v towards the new I.

The nodes may be initialized with random thresholds.

Basic Network Diffusion Models
(source: Kempe et al, KDD'03)

Cascade models are inspired by the dynamics in systems of interacting particles.

Independent Cascade Model:

- Starting with an initial set of active nodes A_0
- at iteration j
  - each *newly activated* node w (w became active at j-1) gets the chance

    to activate each inactive neighbour v

    and succeeds with likelihood $p_{w,v}$
- until no new activations take place.

Influence Maximization
Different formulations

Given is a network.

We want to choose a set of nodes, from which the influence will spread across the network.

- What is the minimal set of nodes to choose, so that the whole network is activated?
- For a given number k, which k nodes should we choose so that a maximal subset of the network is activated?
- The motivation of a node incurs a node-dependent cost. For a given budget B, which set of nodes should we choose so that a maximal subset of the network is activated?

Influence Maximization
Different formulations

Given is a network.

We want to choose a set of nodes, from which the influence will spread across the network.

- What is the minimal set of nodes to choose, so that the whole network is activated?

Domingos & Richardson

- For a given number k, which k nodes should we choose so that a maximal subset of the network is activated?
- The motivation of a node incurs a node-dependent cost. For a given budget B, which set of nodes should we choose so that a maximal subset of the network is activated?

Recall: Spread of influence in a network
Problem formalization and analysis in (Kempe et al, KDD'03)

We observe a Social Network as a medium for the spread of an idea, innovation, item I:

- Understand the network diffusion processes for the adoption of the new I.

  Well-studied problem in social sciences, among else for the acceptance of medi innovations

- Given is a network N.
  We want to promote a new I to that set S of individuals, such that a maximal set of further adoptions will follow.

  "Influence Maximization Problem"
  New formal problem p posed by Domingos and Richardson

Spread of Influence in a Network
The contribution of (Kempe et al, KDD'03) – 1 of 4

In their KDD'03 paper

Maximizing the Spread of Influence through a Social Network

David Kempe, Jon Kleinberg and Eva Tardos:

- formulate the Influence Maximization Problem as a new problem ***p***
- position ***p*** into the theory of diffusion models, which have been widely studied in the social sciences
- prove that ***p*** is NP-hard
- show that the linear threshold model and the independent cascade model deliver solutions that are within 63% (1-1/e) of the optimal for ***p***

Spread of Influence in a Network
The contribution of (Kempe et al, KDD'03) – 2 of 4

In their KDD'03 paper

Maximizing the Spread of Influence through a Social Network

David Kempe, Jon Kleinberg and Eva Tardos:

- formulate the Influence Maximization Problem as a new problem ***p***
- propose a category of models for ***p*** by selecting influence functions from the family of *submodular functions*
- prove that this whole category of models achieves solutions within 63% of the optimal

An Impact-Oriented View upon Communities

- Tracing the influential members in a group of individuals
- Patterns of influence in a social network
- Being influenced to join a community

Patterns of influence in social networks

Individuals that have a central position in a network have the potential to influence their neighbours.

- What do influence patterns look like?
  - Stars => Only one level of influence => no proliferation
  - Trees => Opinions, ideas, information coming from an influential individuum is taken over and spread across the network
  - Graphs with nodes having high in-degree => Nodes that receive, combine (and possibly spread) influence from multiple individuals
  - Circles
- How is influence proliferating in a network?

"Cascades" in a recommendation network
The method of (Leskovec, Singh & Kleinberg, PAKDD'06)

... Information cascades are phenomena in which an action or idea becomes widely adopted due to influence by others. .. (Leskovec, Singh & Kleinberg, PAKDD'06)

"Cascades" in a recommendation network
The method of (Leskovec, Singh & Kleinberg, PAKDD'06)

... Information cascades are phenomena in which an action or idea becomes widely $\boxed{\text{adopted}}$ due to $\boxed{\text{influence}}$ by others. .. (Leskovec, Singh & Kleinberg, PAKDD'06)

An information cascade is more than information dissemination.

A cascade is a pattern of influence.

"Cascades" in a recommendation network
The method of (Leskovec, Singh & Kleinberg, PAKDD'06)

... Information cascades are phenomena in which an action or idea becomes widely adopted due to influence by others. .. (Leskovec, Singh & Kleinberg, PAKDD'06)

Objectives:

- Modeling *influence* in a recommendation network
- Discovering patterns of influence – cascades
- Understanding the structure of cascades
  - Are they stars around a center, trees that reflect a spread of influence, or are they more complex?
  - What is the interplay between the underlying network and the cascades we see in it?

The dataset of the recommendation network
(Leskovec et al, ACM TOW 2007)

- Dataset
  - ~ 4 million people
  - ~ 16 million recommendations on
  - ~ 500,000 products
  - Collected from June 2001 to May 2003

Group	Number of nodes Purchases	Forward	Percent
Book	65,391	15,769	24.2
DVD	16,459	7,336	44.6
Music	7,843	1,824	23.3
Video	909	250	27.6
Total	90,602	25,179	27.8

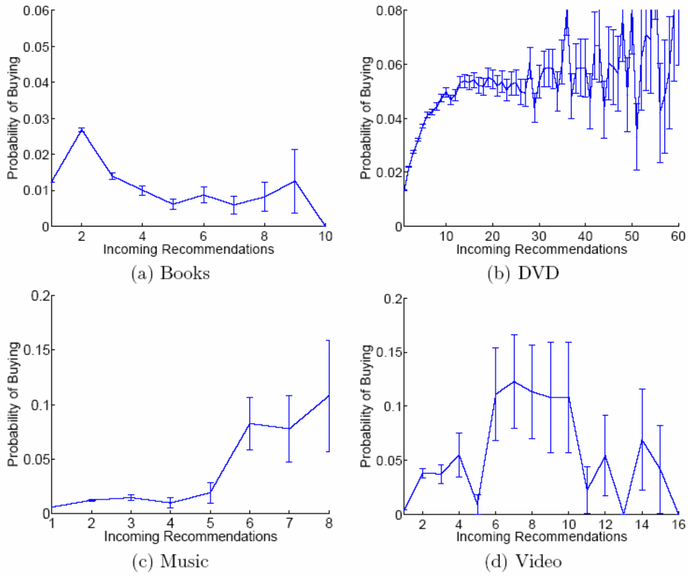
Group	p	n	r	e	b_b	b_e
Book	103,161	2,863,977	5,741,611	2,097,809	65,344	17,769
DVD	19,829	805,285	8,180,393	962,341	17,232	58,189
Music	393,598	794,148	1,443,847	585,738	7,837	2,739
Video	26,131	239,583	280,270	160,683	909	467
Full network	542,719	3,943,084	15,646,121	3,153,676	91,322	79,164

The method of (Leskovec, Singh & Kleinberg, PAKDD'06)
Modeling for the recommendation network

The model was designed with the specific network in mind:

- An individual can perform two actions of relevance:
  - purchase a product
  - recommend a purchased product to another individual at the timepoint of purchase
- The graph is temporal in nature:
  - Node:= individual
  - Edge (source,target,p,t) :=
    The source recommended product p to target at timepoint t
- There is an incentive in recommending products:
  - The *first* node that launches a recommendation leading to a purchase gets a discount.

Success of Recommendations in the network (Leskovec et al, ACM TOW 2007)



Probability of buying given a number of incoming recommendations

The method of (Leskovec, Singh & Kleinberg, PAKDD'06) Challenges and assumption in modeling cascades

Challenges posed by the specific network:

- Events that complicate the analysis:
  - An individual may receive recommendations after having purchased a product.
  - An individual may purchase the same product many times.
- Assumption:
  - If a node receives a recommendation, buys the product and recommends it later on, then we have a ***cascade***.

ATTENTION:

- A person has *no* incentive to recommend a product already recommended to him/her.

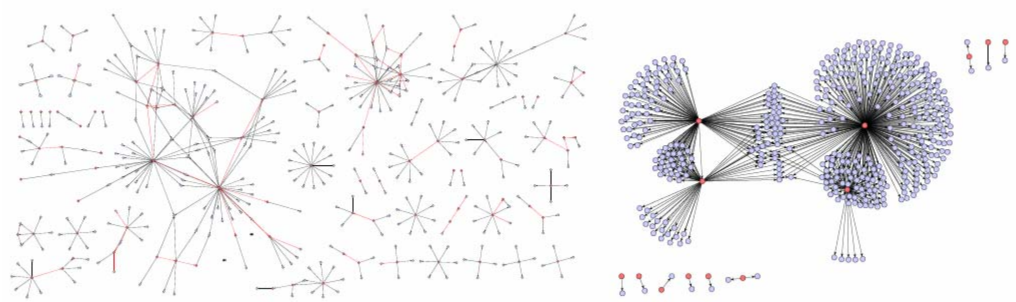
The method of (Leskovec, Singh & Kleinberg, PAKDD'06)
Cleaning the graph and mining cascades

- Cleaning the graph:
  - Recommendations that did not lead to a purchase were eliminated.
  - Recommendations that were delivered after the purchase were eliminated.
- Enumerating *local cascades*:
  - For each node *v*, only edges up to *h* hops away are considered (independently of direction).
- Subgraph matching:
  - Small cascades are matched exactly (allowing for isomorphisms).
  - Large cascades are matched approximately on their *signatures.*
    - A signature encompasses number of nodes, number of edges, in-degree and out-degree of nodes.

The method of (Leskovec, Singh & Kleinberg, PAKDD'06)
Findings for four product categories

- Size distribution of cascades
  - All cascades follow power-laws.
  - Products of one category (DVDs) show a significantly different distribution – many large cascades.
- Structure of frequent cascades
  - The majority of cascades is simple.
  - Many cascades are one-level trees (stars), while
  - there are also cascades with common recipients of recommendations.
  - The DVD product category exhibits larger and denser cascades.

Structures in the recommendation network (Leskovec et al, ACM TOW 2007)



(a) Medical book (b) Japanese graphic novel

Two examples:

(a) First aid study guide *First Aid for the USMLE Step*,

(b) Japanese graphic novel (manga) *Oh My Goddess!: Mara Strikes Back*.

Rewind on the method of (Leskovec, Singh & Kleinberg, PAKDD'06)

- A case-driven contribution, using simple graph matching algorithms and a reasonable model of influence cascades
- and delivering insights for a very large recommendation network.
  - Disregarding the incentive system of the network, there are many cascades, remarkably dense in one product category.
- What about ...
  - the role of the incentive system?
  - the differences among the product categories?
  - communities?

An Impact-Oriented View upon Communities

- Tracing the influential members in a group of individuals
- Patterns of influence in a social network
- Being influenced to join a community

Influence and community evolution

- What moves an individuum to join a community?
  - Understanding the role of influential members on the participation decision
  - Understanding the patterns of proliferating influence
- How does a community evolve with respect to its members?
  - Modeling and tracing evolving communities
  - Modeling the dynamic aspects of communities

BLOCK 3: Community Dynamics

References for
Block 2 - An Impact-Oriented View upon Communities

- P. Domingos, M. Richardson "Mining the Network Value of Customers", Proc. of KDD'01, p. 57-66
- D. Kempe, J. Kleinberg, E. Tardos "Maximizing the Spread of Influence through a Social Network", Proc. of KDD'03, p. 137-146
- J. Leskovec, A. Singh, J. Kleinberg "Patterns of Influence in a Recommendation Network", Proc. of PAKDD'06
- J. Leskovec, L.A. Adamic, B.A. Huberman. *The Dynamics of Viral Marketing*, ACM Trans. on the Web, (1)1, 2007

Block 2 is over ...

Thank you!

Questions?

Presentation Outline

- ☑ Block 1: Community models
- ☑ Block 2: Three perspectives for community discovery
  - ☑ Similarity-based perspective
  - ☑ Interaction-based perspective
  - ☑ Impact-based perspective
- Block 3: Community dynamics
- Block 4: Outlook

Influence and community evolution

- What moves an individuum to join a community?
  - Understanding the role of influential members on the participation decision
  - Understanding the patterns of proliferating influence
- How does a community evolve with respect to its members?
  - Modeling and tracing evolving communities
  - Modeling the dynamic aspects of communities

What moves an individual to join a community?
The influence of network structures (Backstrom et al, KDD'06)

Objectives:

- Identifying structures that influence the decision of individuals in joining the community
- Understanding the evolution of a community and its interplay (overlap of members) with other communities

Backstrom et al study *known communities,*
defined explicitly by their members.

- Application 1: DBLP
  Community := Authors of articles in a given *conference*
- Application 2: Live Journal
  Community:= Declared friends of a person in *Live Journal*

Influence of a community on non-members
(Backstrom et al, KDD'06)

Hypothesis:

- The propensity of an individual to join a given community depends on the number of friends the individual has inside that community.

Modeling a community and its fringe
(Backstrom et al, KDD'06)

Model:

- A community is a subgraph of interacting members.
- A community has a "fringe": It consists of individuals that interact with at least k community members but are not community members themselves.

Approach:

- Identify the features that influence members of the fringe to move inside the community.
  - Number of friends in the community
  - Iintensity of interaction with those friends
  - Intensity of interaction among the community friends, ...

Influence of a community on non-members
(Backstrom et al, KDD'06)

Hypothesis:

- The propensity of an individual to join a given community depends on the number of friends the individual has inside that community.

Findings:

- The likelihood of joining a community increases with the number of friends already in it,
  but is very noisy for individuals with many friends.
- The existence of friendships among friends contributes to this likelihood.
- The two variables make a good predictor of membership propensity.

Influence and community evolution

- What moves an individuum to join a community?
  - Understanding the role of influential members on the participation decision
  - Understanding the patterns of proliferating influence
- How does a community evolve with respect to its members?
  - Modeling and tracing evolving communities
  - Modeling the dynamic aspects of communities

Capturing community evolution on a data stream

Objectives:

- Detect and understand changes on a existing structures of the social network
  - communities that vanish
  - communities that merge or split
- Detect new structures – emerging communities

Basic approach:

- The data stream is captured at timepoints $t_1,...,t_n$.
- At each timepoint t_i, the patterns of the previous timepoint are juxtaposed (?) to the new data.

Mining an evolving graph of interactions
The method of (Aggarwal & Yu, SDM'05)

In "Online Analysis of Community Evolution in Data Streams", Aggarwal and Yu elaborate on the discovery of *expanding*, *contracting* and *stable* communities.

Components of the approach:

- a model for the stream of interactions
- a definition of "evolving community"

  A cluster of interactions that evolves differently from its surroundings
- an algorithm that traces evolving communities
- a measure of a community' evolution

Community dynamics in CODYM
The method of (Falkowski et al, Web Intelligence'06)

Components:

- A mechanism that finds communities upon a frozen part of the data (a time period)
- A method that partitions the horizon of observation in periods
- A model that captures the notion of "community" across time periods
- A mechanism that highlights community dynamics
- Visualization aids to community evolution monitoring

The method of (Falkowski et al, Web Intelligence'06)
Subgroup detection upon a static network

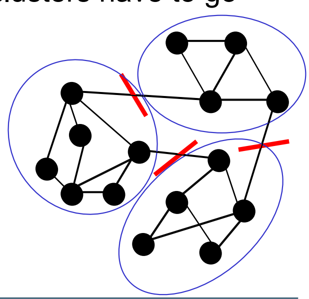
Core idea:

- The network is partitioned into groups with *hierarchical divisive clustering*.
- The partitioning is done by removing edges according to the *edge betweenness* criterion of (Girvan & Newman,2002).
- The output of the clustering algorithm is a dendrogram.
  - It is "cut" at some level.
    The clusters are the graph partitions at this level.
  - The cut is performed according to a quality measure of (Newman & Girvan, 2004).

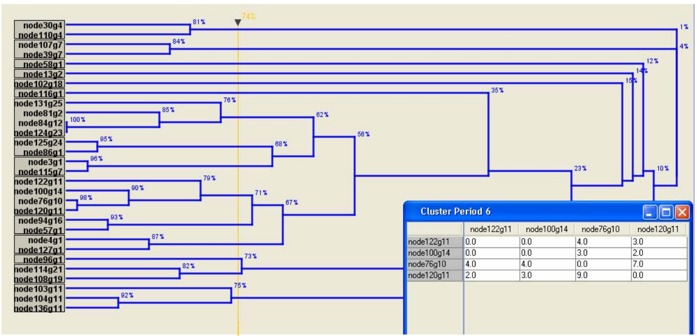
Subgroup detection upon a static network:
Edge Betweeneness in divisive clustering

- Motivation (and assumption):
  - The subgroups/communities are tightly bound clusters, loosely connected to their surroundings.
- The concept (Girvan & Newman, 2002):
  - When a graph is made of tightly bound clusters, loosely interconnected, all shortest paths between clusters have to go through the few intercluster connections.
  - For each edge, we count the number of shortest paths that go through it.

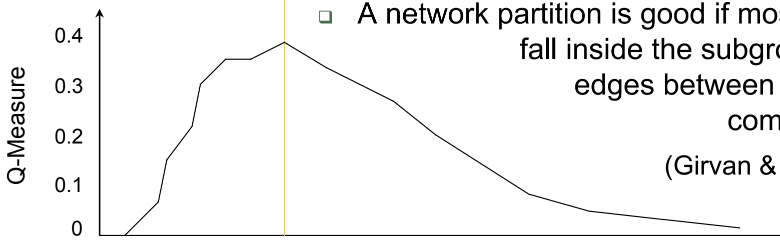
```
repeat until no more edges in graph g
Compute edge betweenness for all edges
Remove edge with highest betweenness
end
```



Subgroup detection upon a static network:
Quality measure for cutting the dendrogram



The Dendrogram



- A network partition is good if most of the edges fall inside the subgroups, while the edges between subgroups are comparatively few.

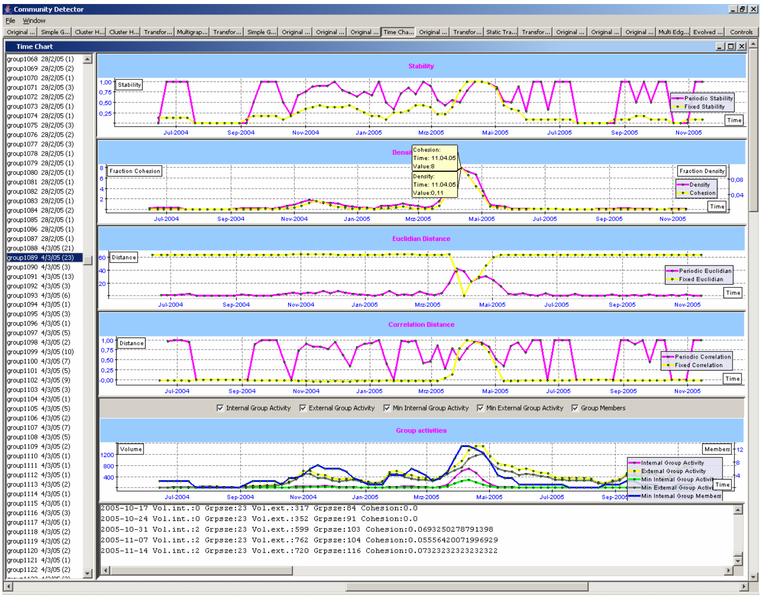
(Girvan & Newman, 2004)

Community dynamics in CODYM
The method of (Falkowski et al, WebIntelligence'06)

Components:

- ☑ A mechanism that finds communities upon a frozen part of the data (a time period)
- A method that partitions the horizon of observation in periods
- A model that captures the notion of "community" across time periods
- A mechanism that highlights community dynamics
- Visualization aids to community evolution monitoring

Studying one subgroup across time:
Visualization of statistical measures at earlier and later time slots



Finding similar subgroups within a window of τ time periods

- Two subgroups are similar if they have many members in common.

Concept:

- For two subgroups X, Y *found in different periods*:

$$\text{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|,|Y|)} \qquad \text{sim}(X,Y) = \begin{cases} 1 & \text{overlap}(X,Y) \geq \tau_{\text{overlap}} \\ 0 & \text{otherwise} \end{cases}$$

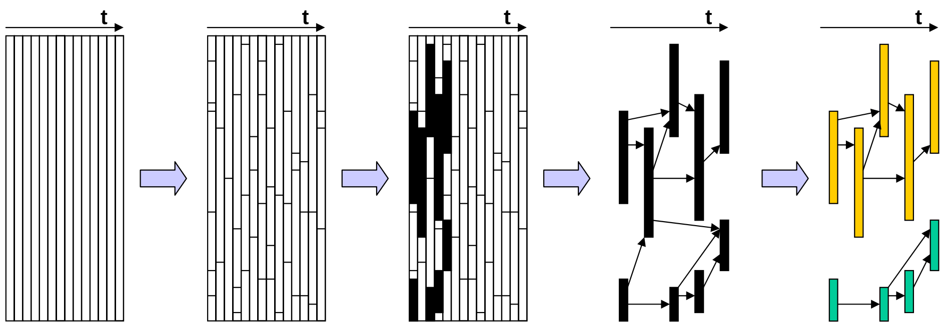
- from which we derive a similarity function subject to the time window τ_{periods}:

$$\text{similarity}(X^{G_i}, Y^{G_j}) = \begin{cases} 1 & |t_i\text{-}t_i| \leq \tau_{periods} \wedge \text{overlap}(X,Y) \geq \tau_{\text{overlap}} \\ 0 & \text{otherwise} \end{cases}$$

The method of (Falkowski et al, Web Intelligence'06)
Subgroup vs. Community

- The new termini:
  - A ***community*** is a cluster of similar subgroups
  - A subgroup found at t_i is a *community instance*
- The approach:
  - Similar subgroups (subject to the time window) are connected with edges
  - The resulting graph is partitioned into clusters with hierarchical divisive clustering
  - The partitioning is done by removing edges according to the edge betweenness criterion
- So, a community is a cluster of subgroups that evolve but still remain tightly bound to each other, maintaining loose connections to other subgroups.

The method of (Falkowski et al, Web Intelligence'06)
Overview



Step 1. Partitioning the time axis

Step 2. First clustering to find subgroups (community instances) in time windows

Step 3. Detecting similar community instances in time windows

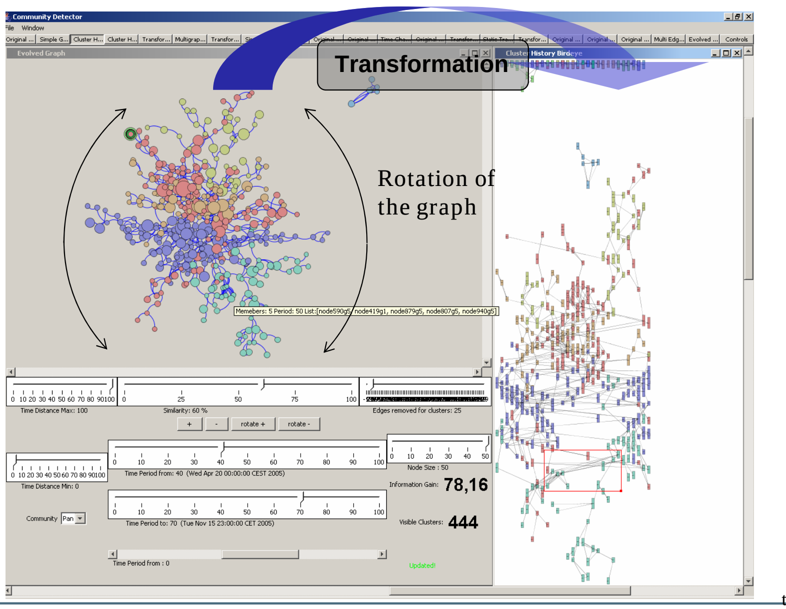
Step 4. Visualization of similar community instances

Step 5. Second clustering to find clusters of similar community instances

Data Set

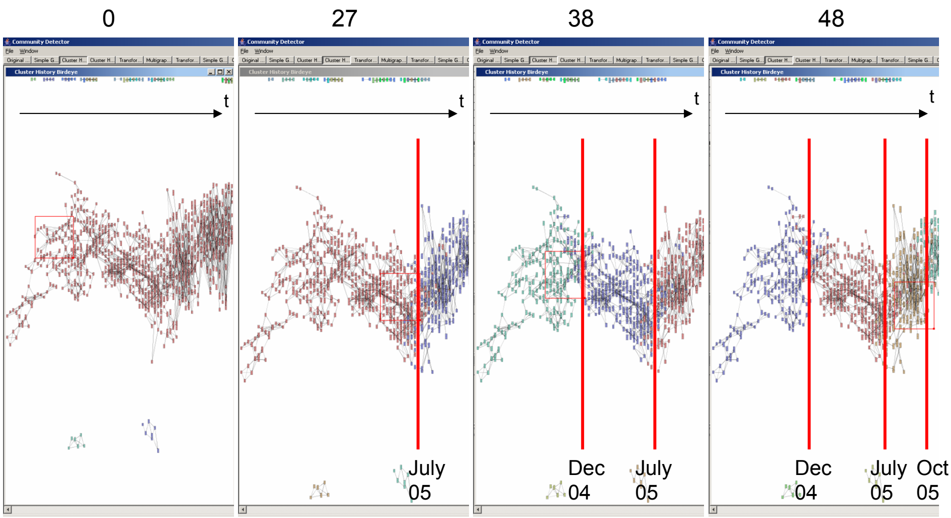
- Data Set
  - approx. 1,000 actors
  - 250,000 interactions (guestbook entries)
  - over a period of 18 months (June 2004 – November 2005, 75 weeks)
- Sliding Window Approach
  - Window length of 14 days; step width of ½ of the window length

Visualization:Community Instances & Communities



Building and visualizing communities: Experiments on a site of guest & foreign students

Number of clustering iterations (= number of edges removed):



Community dynamics in CODYM
The method of (Falkowski et al, WebIntelligence'06)

Components:

- ☑ A mechanism that finds communities upon a frozen part of the data (a time period)
- A method that partitions the horizon of observation in periods
- ☑ A model that captures the notion of "community" across time periods
- ☑ A mechanism that highlights community dynamics
- ☑ Visualization aids to community evolution monitoring

References for Block 3: Community Dynamics

- L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan "Group Formation in Large Social Networks: Membership, Growth and Evolution", Proc. of KDD'06, p. 44-54
- Charu Aggarwal and Philip Yu "Online Analysis of Community Evolution in Data Streams", Proc. of SIAM Data Mining Conf., 2005.
- T. Falkowski, J. Bartelheimer, M. Spiliopoulou "Mining and Visualizing the Evolution of Subgroups in Social Networks", Proc. of IEEE/WIC/ACM Int. Conf. on Web Intelligence (WI'06), Hong Kong, Dec. 2006

Presentation Outline

- ☑ Block 1: Community models
- ☑ Block 2: Three perspectives for community discovery
  - ☑ Similarity-based perspective
  - ☑ Interaction-based perspective
  - ☑ Impact-based perspective
- ☑ Block 3: Community dynamics
- Block 4: Outlook

Summarizing the landscape

Communities are modeled and studied from different perspectives.

- Data mining is applied, among else, to:
  - discover communities, i.e. groups of instances that adhere to an a priori defined model
    - persons with similar interests
    - persons that navigate in a similar way
    - persons that interact
    - persons that influence each other
  - derive recommendations for a person on the basis of
    - people most similar to her
    - people with similar interests and preferences
    - people of potential influence upon her (including people she trusts)
  - study the dynamics of communities
    - to understand how communities emerge, evolve and stagnate
    - to gain insights on the role of individuals in a community

Active user community discovery

- Discovery of Web user communities.
  - Analysis of usage data.
  - Discovery of interest and navigation patterns.
  - Communities of content consumers.
- Discovery of Web communities.
  - Analysis of Web structure.
  - Discovery of graph patterns (linkage of pages).
  - Communities of content creators.

Active user community discovery

- Web users are increasingly becoming content creators and service providers.
- At the same time they remain content consumers and service users.
- Many new services support active users:
  - Users as publishers, e.g. blogs, fora etc.
  - Collaborative creation of content and knowledge, e.g. flickr, del.icio.us, Yahoo!Answers, Wikipedia, bibsonomy, etc.

Active user community discovery

- Active user community discovery combines the existing approaches, taking into account:
  - Usage: what the user has chosen to “consume“.
  - Content: what the user has contributed
  - Structure: links between content created by different users.
- Additionally it introduces a range of new issues:
  - Consumption-creation pattern discovery.
  - Separating characteristics between consumer and creator sub-communities.
- Active user community models combine this information into comprehensive generic user models.
- Discovery can also help evolve (manually created) communities.

Community and environs

Communities are at the visier of malefactors.

- How to protect a community from spam content?
- How to secure community property (including shared intellectual property and person-private information) against adversaries?

Different types of solutions:

- Spam detection
- Security measures against intruders
- Privacy-preserving measures against adversaries
- Reputation mechanisms
- Communities of trust

A few words on trust and reputation

Communities of Trust

- **Figallo states:**
  "Trust is the social lubricant that makes community possible.", in Figallo, Cliff. *Hosting Web Communities* (New York: John Wiley & Sons, Inc.) 1998
- **Trust**: Community members know with whom they 're dealing and that it's safe to do so.
- Without trust a community cannot function.
- Trust is basis for *reputation*. Key elements are:
  - Letting members build trust over time.
  - Posting clear policies regarding privacy and online actions and abiding by them.
  - Allowing different levels of privacy so members can reveal more about themselves as they get to know each other.
  - Providing experts with certifications and detailed profiles so members are able to trust that "experts" have the qualifications they claim.
  - Allowing member verification of profiles.
  - Hands-off management that garners more trust and encourages greater self-governance than interfering or policing management.

Communities of Reputation

- ***Reputation***: Reputation is what is generally said or believed about a person's or thing's character or standing. *(Concise Oxford Dictionary)*
- Reputation vs. Trust:
  - "I trust you because of your good reputation."
  - "I trust you despite your bad reputation."
- Trust is a personal and subjective phenomenon
- Reputation is a collective measure of trustworthiness
- Reputation lies at the juncture between identity and trust and influences behavior in several ways.
- Reputation measures give members a way to evaluate each other, so they know whom to trust, or whom not to trust.
- It helps people form the best alliances to get the desired information; and the desire to have a good reputation discourages bad behavior and encourages members to request feedback

Reputation Network Architectures

- Centralized Reputation Systems
  - A "reputation center" collects ratings for a given community member from other community members who know him.
  - The reputation centre derives a reputation score for every participant, and makes all scores publicly available.
  - The idea is that transactions with reputable participants are likely to result in more favorable outcomes than transactions with disreputable participants.
- Distributed Reputation Systems
  - Distributed reputation stores instead of a single center.
  - Ratings are submitted when members are interacting with each other.
  - A community member who wants to interact with another member, must find the distributed stores or obtain ratings from as many community members as possible who have had interaction experience with the examined member.

Reputation Metrics

- Simple Summation or Average of Ratings
  - Sum the number of positive ratings and negative ratings separately, and keep a total score as the positive score minus the negative score.
- Bayesian Systems
  - Input: binary ratings (pos, neg)
  - Output: a-posteriori reputation score, based on the a-priori score and the new ratings
  - Reputation score: beta probability density function (PDF):

$$beta(p \mid a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1}(1-p)^{b-1}$$

  - a,b represent the amount of positive and binary ratings

Reputation Metrics

- Discrete Trust Models
  - Use discrete statements not continuous measures, e.g. trustworthiness x can be referred as *Very Trustworthy*, *Trustworthy*, *Untrustworthy* and *Very Untrustworthy*.
- Flow Models
  - A participant's reputation increases as a function of incoming flow, and decreases as a function of outgoing flow. (e.g. PageRank)

Trust & Reputation Systems

System	Trust & Reputation Mechanism
GroupLens	rating of articles
OnSale	buyers rate sellers
Epinions	number of reviews
Firefly	rating of recommendations
EBay	buyers rate sellers

Discovering and Tracking User Communities

Thank you!

Questions?