Community Detection: From Plain to Attributed Complex Networks

Martin Atzmueller

University of Kassel, Research Center for Information System Design
Ubiquitous Data Mining Team, Chair for Knowledge and Data Engineering

Web Science 2016, Hannover – 2016-05-22

Exploratory Data Analysis

- Different aspects & perspectives
  - Hypothesis generating
  - Visualization & Analytics
  - Semi-automatic & Interactive
  - Detect local models
- Approaches & methods
  - Local exceptionality detection
  - Community detection
  - Description-oriented community detection

Pattern

- Merriam Webster: "A repeated form or design especially that is used to decorate something"

- Oxford: "An arrangement or design regularly found in comparable objects"
- Pattern in data mining [Bringmann et al. 2011]
  - Captures regularity in the data
  - Describes part of the data

Attributed Graphs

- Additional information (on nodes, edges)
- E.g., "knowledge graph"

Homophily (i.e. "Love of the same")

- Sociology:"Birds of a feather flock together" (Lazarsfield & Merton 1954)
- Social Networks: "Similarity breeds connection": A connection between similar people occurs at a higher rate than between dissimilar ones. (Mc Pherson et al. 2001)

Attributed Network/Graph

II Networks in the real world

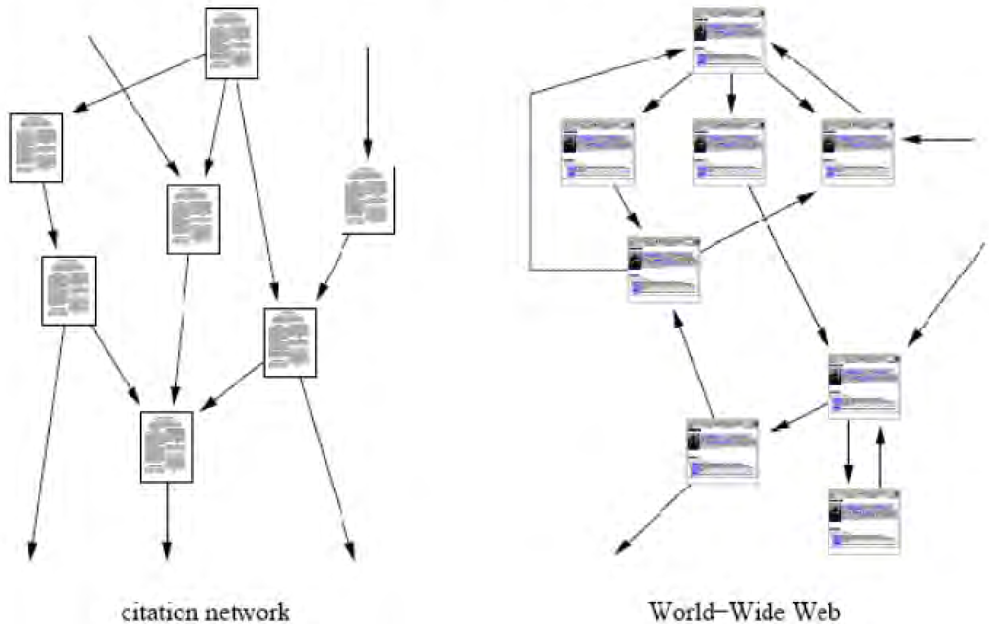
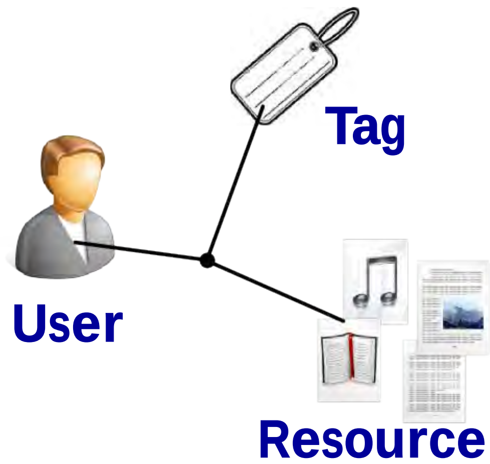


FIG. 4 The two best studied information networks. Left: the citation network of academic papers in which the vertices are papers and the directed edges are citations of one paper by another. Since papers can only cite those that came before them (lower down in the figure) the graph is acyclic—it has no closed loops. Right: the World Wide Web, a network of text pages accessible over the Internet, in which the vertices are pages and the directed edges are hyperlinks. There are no constraints on the Web that forbid cycles and hence it is in general cyclic.

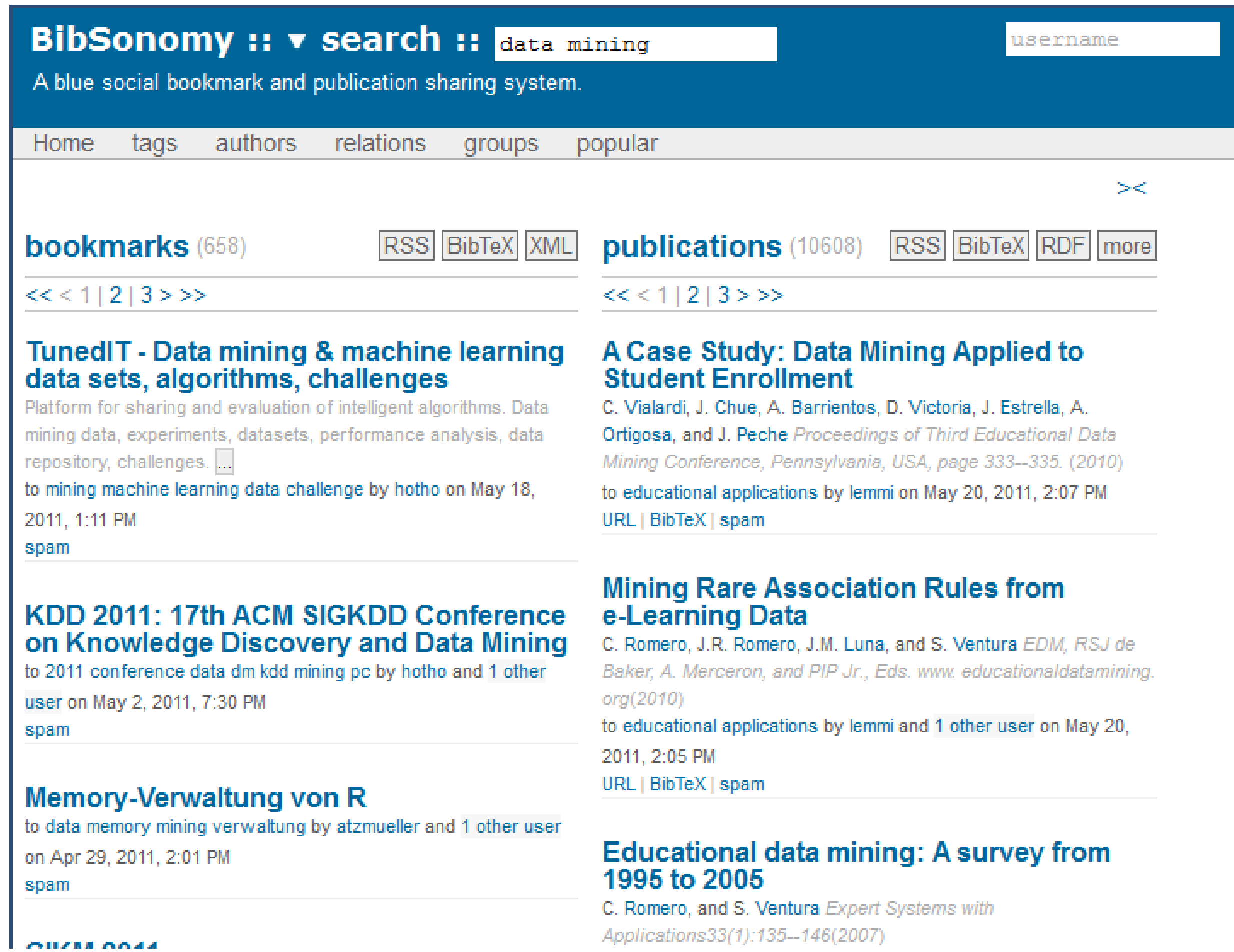
(Newman 2003)

- Examples
  - Citation Attributes
    - (Co-)Authors
    - Affiliation
    - Country
    - Gender
    - …
  - WWW
    - Links
    - Content (BoW)
    - …

Real-World System I: BibSonomy



http://www.bibsonomy.org



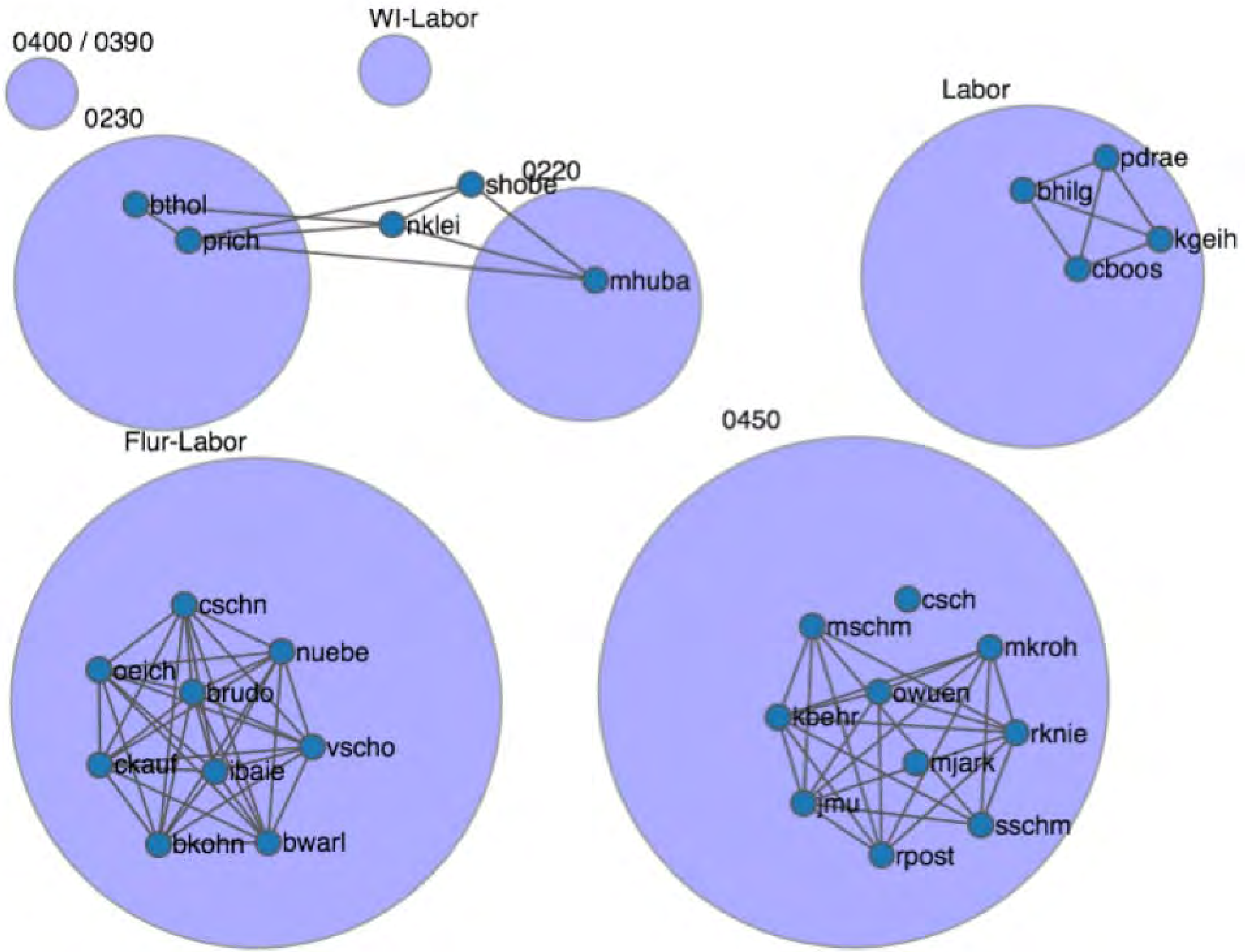
- Users assign tags to resources
  - → Organize
  - → Share
  - → Categorize

Real-World System II: Conferator

- Social Conference Guidance System
  - GI: Lernen – Wissen – Adaptivität (LWA) 2010 + 2011 + 2012
  - ACM Hypertext 2011
  - INFORMATIK 2013
  - UIS 2015
- Based on RFID-Technology (smart badges)
- Management of social contacts, personalization of conference schedule
- Localization

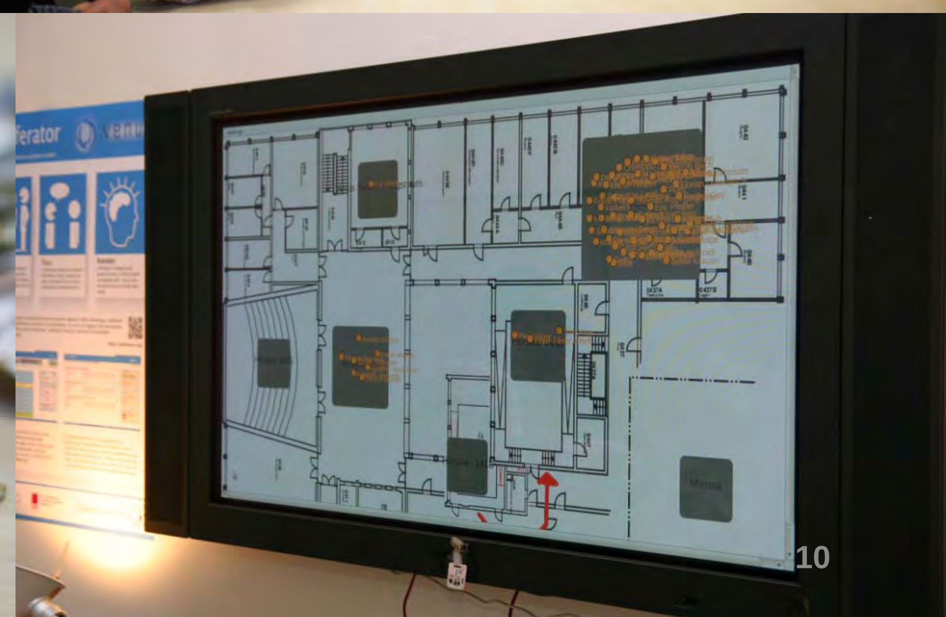
www.conferator.org

Conferator - Live Interaction

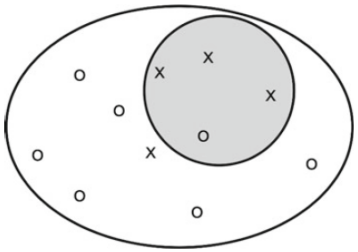


Conferator

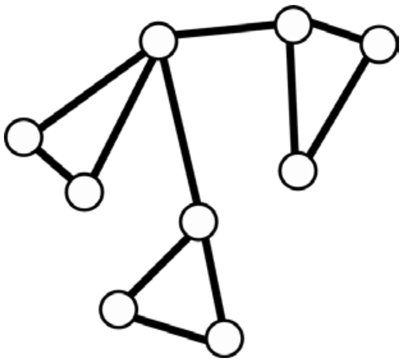
- Social interaction networks:
  - Friend network
  - Contact network
  - Picked/Visited talks
  - Co-location network

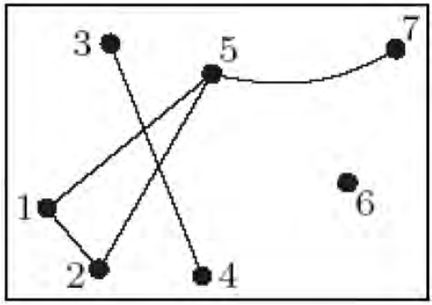


Agenda



- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook





Terminology

Network ➔ Graphs

- Set of atomic entities (actors)
  ➔nodes, vertices
- Set of links/edges between nodes ("ties")
- Edges model pairwise relationships
- Edges: Directed or undirected
- Social network [Wassermann & Faust 1994]
  - Social structure capturing actor relations
  - Actors, links given by dyadic ties between actors (friendship, kinship, organizational position, …)
    ➔ Set of nodes and edges
  - Abstract object – independent of representation

Variables [Wassermann & Faust 1994]

- Structural
  - Measure ties between actors (➔ links)
  - Specific relation
  - Make up **connections** in graph/network
- Compositional
  - Measure actor attributes
    - Age
    - Gender
    - Ethnicity
    - Affiliation
    - …
  - **Describe** actors

Attributed Graphs

- Graph: edge attributes and/or node attributes
  - Structure: ties/links (of respective relations)
- Attributes - additional information
  - Actor attributes (node labels)
  - Link attributes (information about connections)
  - Attribute vectors for actors and/or links
  - … can be mapped from/to each other
- Integration of heterogenous data (networks + vectors)
- Enables simultaneous analysis of relational + attribute data

Subgroups & Cohesive subgroups

[Wasserman & Faust 1994]

- Subgroup
  - Subset of actors (and all their ties)
- Define subgroups using specific criteria (homogeneity among members)
  - Compositional – actor attributes
  - Structural – using tie structures
- Detection of cohesive subgroups & communities ➔ structural aspects
- Subgroup discovery ➔ actor attributes
- … attributed graph ➔ can combine both

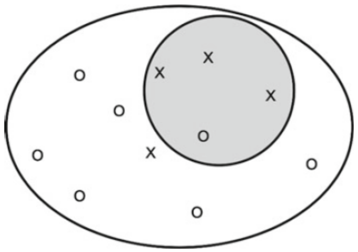
Cohesive Subgroups [Wasserman & Faust 1994]

- Components: Simple, detect "isolated" islands
- Based on (complete) mutuality
  - Cliques
  - n-Cliques
  - Quasi-cliques
- Based on nodal degree
  - K-plex
  - K-core

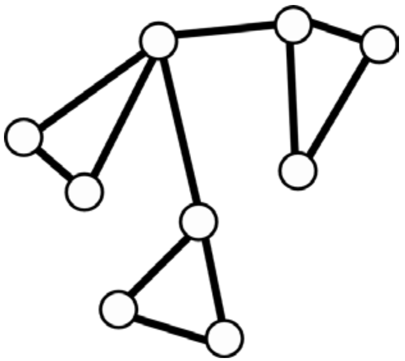
Compositional Subgroups

- Detect subgroups according to specific compositional criteria
  - Focus on actor attributes
  - Describe actor subset using attributes
- Often hypothesis-driven approaches: Test specific attribute combinations
- In contrast: Subgroup discovery [Atzmueller 2015]
  - Hypothesis-generating approach
  - Exploratory data mining method
  - Local exceptionality detection

Agenda



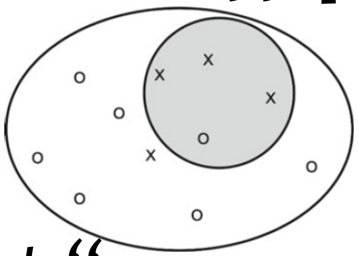
- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook



Subgroup Discovery & Analytics

[Kloesgen 1996, Wrobel 1997]

- Task:
  „Find *descriptions* of subsets in the data, that *differ* significantly for the total population with respect to a *target concept.*“



- Examples:
  - "45% of all men aged between 35 and 45 have a **high income** in contrast to only 20% in total."
  - "66% all all woman aged between 50 and 60 have a **high centrality** value in the corporate network"
- Descriptive patterns for subgroup
  - Gender= Female ∧ Age = [50; 60] ➔ Centrality = high
  - {flickr, delicious}, {library, android}, {php, web} ➔ Centrality = high

Subgroup Discovery

- Given – INPUT:
  - Data as set of cases (records) in tabular form
  - Target concept (e.g. „high centrality“)
  - Quality function (interesting measure)
- OUTPUT - Result: Set of the best *k* **Subgroups:**
  - Description, e.g., sex=female ∧ age= 50-60
    ➔ Conjunction of *selectors*
  - Size *n*, e.g., in 180 of 1000 cases
  - Deviation
    (p = 60% in the subgroup vs. p_0=10% in all cases)

➔"*Quality*" of the subgroup: weight size and deviation

Subgroup Quality Functions [Atzmueller 2015]

- Consider size and deviation in the target concept

_a_: weight size against deviation (parameter)

$$n^a \cdot (p - p_0)$$

***n:*Size of subgroup (number of cases)**

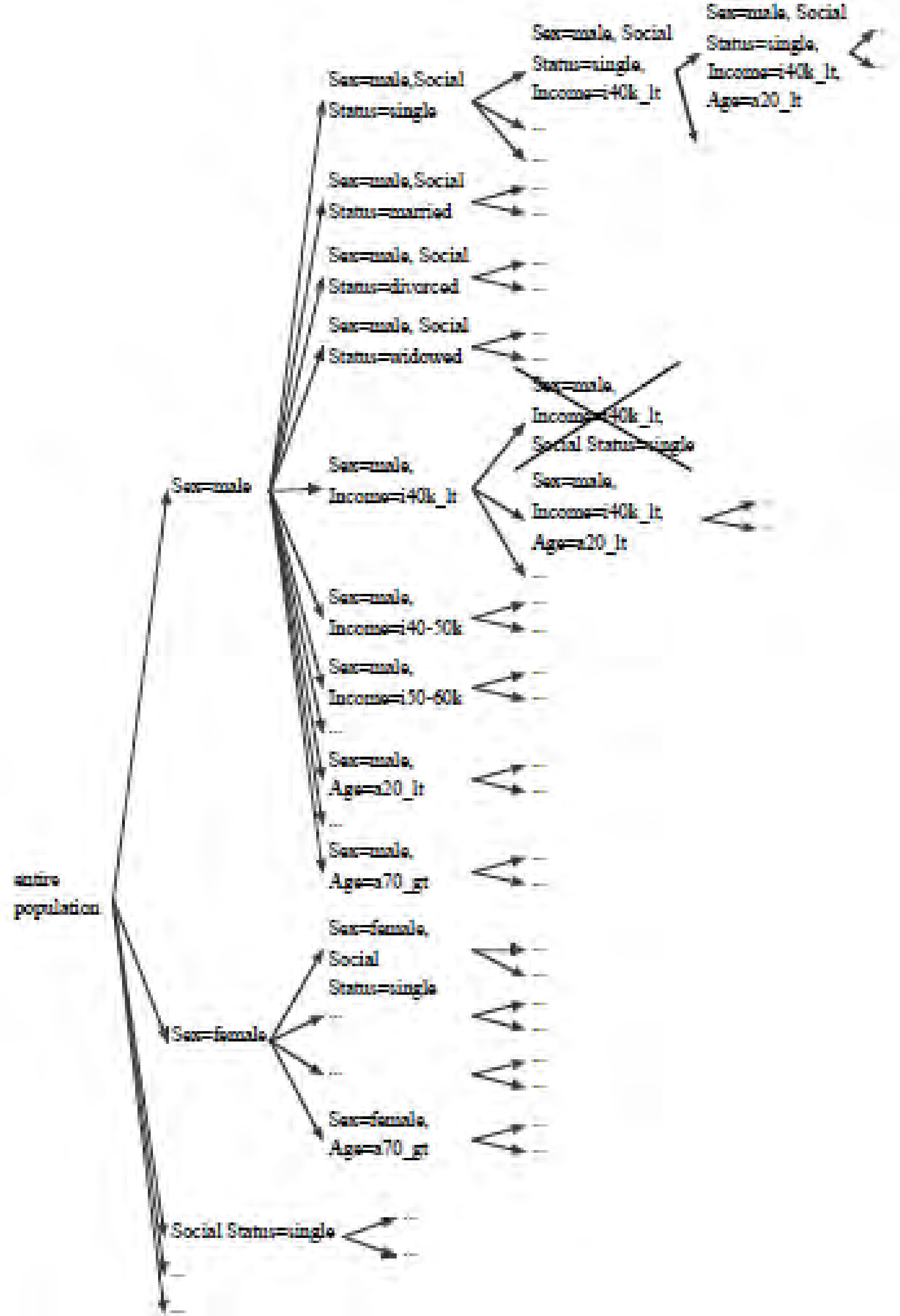
p: share of cases with _target_ = _true_ in the subgroup
p_0: share of cases with _target_ = _true_ in the total population

  - Weighted Relative Accuracy (a = 1)
  - Simple Binomial (a = 0.5)
  - Added Value (a = 0)
- Continous: Mean value (m, m_0) of target variable

$$q_{CWRACC} = \frac{n}{N} \cdot (m - m_0)\,, \quad q_{CPS} = n \cdot (m - m_0)$$

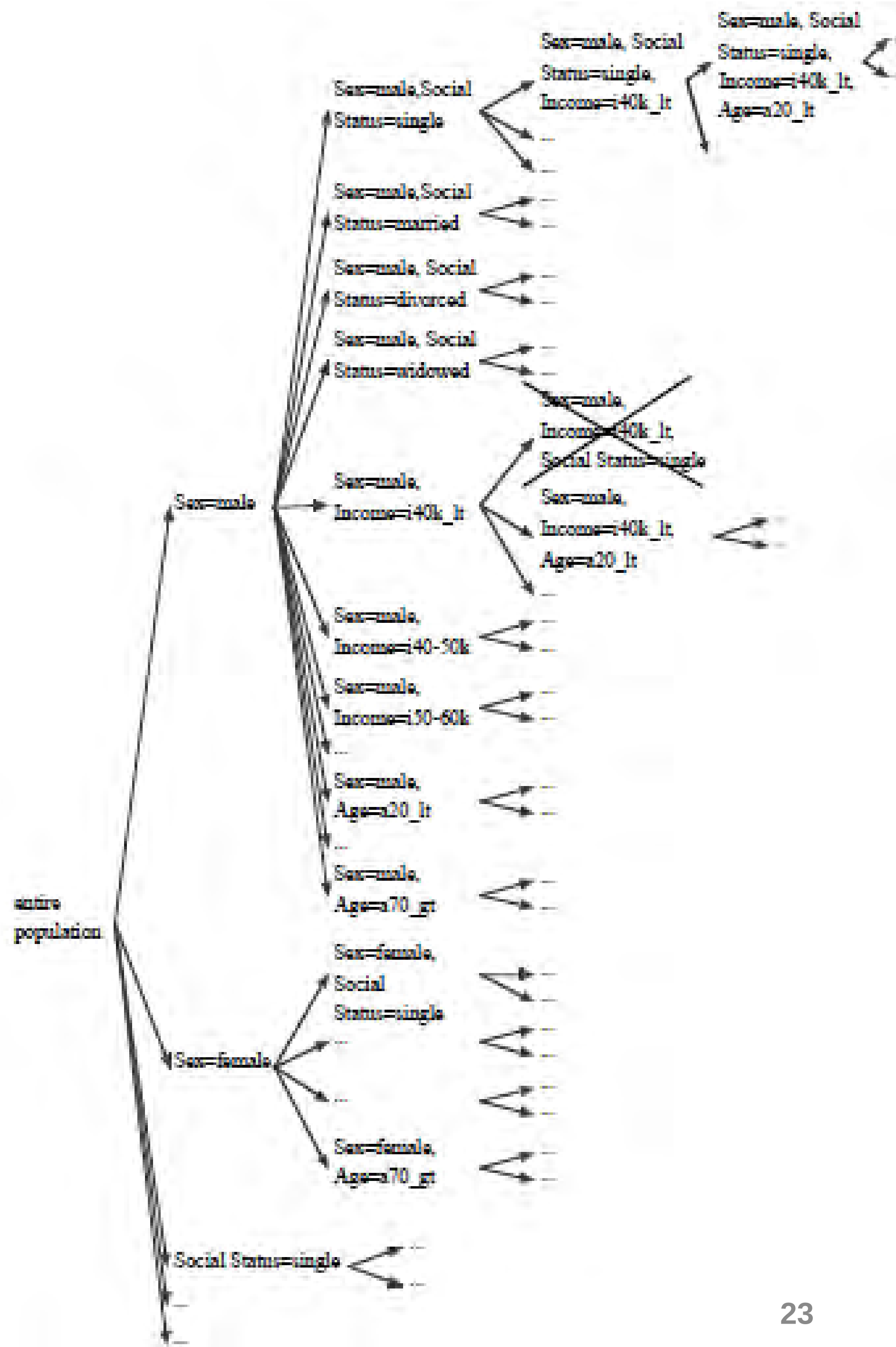
Efficient Search

- Heuristic: Beam Search
- Exhaustive Approaches:
  - Basic idea: Efficient data structures + pruning
  - SD-Map – based on FP-Growth [Atzmueller & Puppe 2006]
  - SD-Map* – Utilizing optimistic estimates (branch & bound) [Atzmueller & Lemmerich 2009]



Pruning

- Optimistic Estimate Pruning – Branch & Bound
- Optimistic Estimate: Upper bound for the quality of a pattern and all its specializations
  ➔Top-K Pruning
- Remove path starting at current pattern, if optimistic estimate for current pattern (and all its specializations) is below quality of worst result of top-k results



Extensions

- Numeric features
- Very large data
  - Distributed Algorithms:
    Local (several cores) vs. network
  - Sampling
- Non tabular data
  - Text
  - Sequences
  - Networks/Graphs (➔ community detection)

Example: Binary target

Income	Sex	Age	Education level	Married	Has Chidren
High	M	>50	High	Y	Y
High	M	>50	Medium	Y	Y
High	F	40-50	Medium	Y	Y
Medium	M	>50	High	Y	N
Medium	M	30-40	Medium	Y	Y
High	M	40-50	Low	N	Y
Low	M	<30	High	Y	N
Medium	F	<30	Medium	Y	N
Low	F	40-50	Low	Y	N
Low	M	40-50	Medium	N	N
Medium	F	>50	Medium	N	N
Low	F	<30	Low	N	N
Low	F	30-40	Medium	N	N
Low	F	40-50	Low	N	N
Low	M	<30	Low	N	N
Medium	F	30-40	Medium	N	N

Target concept: ‚Income' = ‚High'
Quality function: $q = n * (p - p_0)$
$N = 16$; $p_0 = 0.25$

SG 1: ‚Married' = ‚Y'
$n = 8$; $p = 0.375$ ➔ $q = 0.0625$

SG 2: ‚Sex' = ‚M'∧ Age = ‚ < 30'
$n = 2$; $p = 0$ ➔ $q = - 0.03125$

Numeric Features

- Discretization:

  "While only 20% of the total population have an income > 60.000, in subgroup X it can be observed in more than 45% of all cases."

- Mean-Value:
  "While the average income in the total population is 45.000, it is more than 65.000 in subgroup Y. "

➔ Both can be useful,

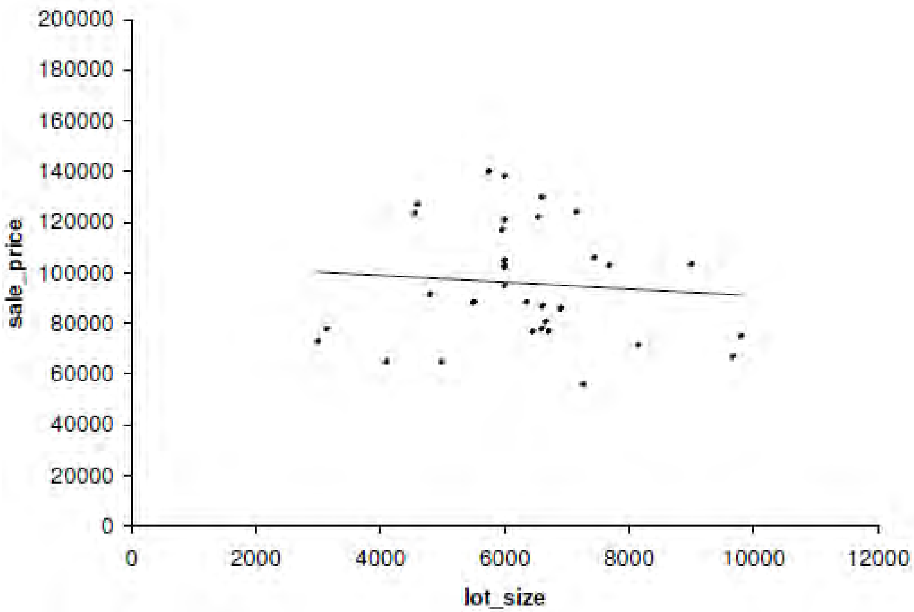
  Mean does not require threshold,

  Is it easier to understand?

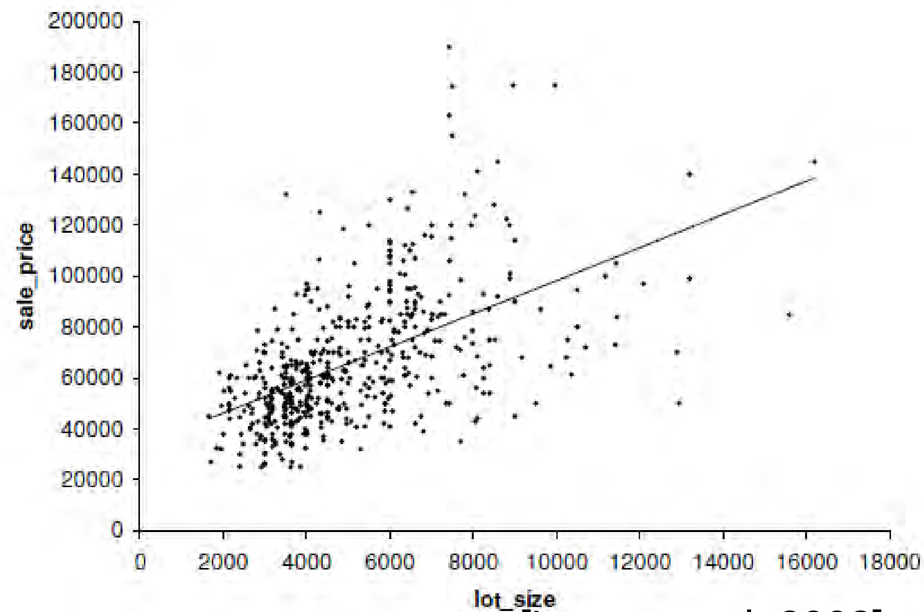
Local Exceptionality Detection

- Exceptional Model Mining
  - Identification of Patterns
  - showing an "interesting behavior" for a certain "model"
    - Mean test (e.g., influence factors for increased centrality)
    - Linear regression (e.g., different centrality measures)
    - Correlation Coefficient (e.g., factors for role analysis)
    - Variance (e.g., degree, clustering coefficient, …)
    - …
- Algorithms:
  - Beam-Search: Heuristic (!) [Duivestein et al. 2015]
  - GP-Growth [Lemmerich et al. 2012]
    - Faster by multiple orders of magnitude compared to standard methods
    - Fastest exhaustive algorithm so far

EMM - Example Linear Regression



Subgroup:
drive = 1 ∧ nbath > 2



[Leman et al. 2008]

Total population

➔ Exploratory Analytics

- Semi-automatic & Interactive
- Hypothesis generating
- Detect local models for description & prediction
  - Local exceptionality detection
    - Subgroup discovery
    - Exceptional model mining
    - … community detection …
  - Applicable also for big data (with Map/Reduce, …)

Subgroup & Pattern Analytics

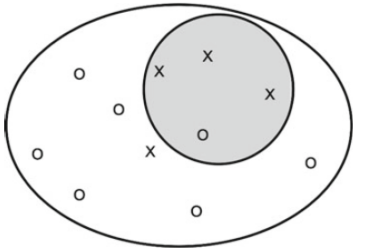
- VIKAMINE: Open-source tools for pattern mining and subgroup analytics

www.vikamine.org

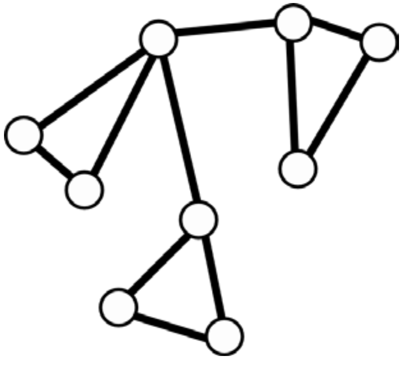
- R package: Algorithms of VIKAMINE

www.rsubgroup.org

Agenda



- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook



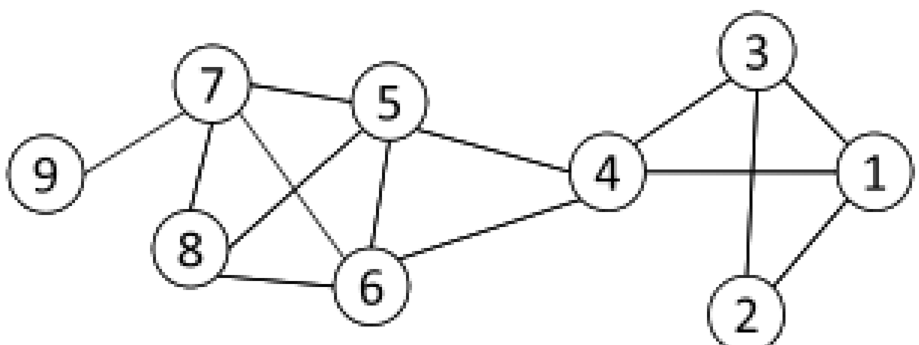
Cohesive Subgroups

- Identify cohesive subgroups of actors
- Cohesive subgroup
  (Wassermann & Faust, p. 249):
  - Subsets of actors
  - Relatively strong, direct, intense , frequent or positive ties
- Social cohesion – primary criterion based on internal ties
- Extension: Social structure
  (➔ communities!)

Subgroups – Local Definitions

[Wasserman & Faust 1994]

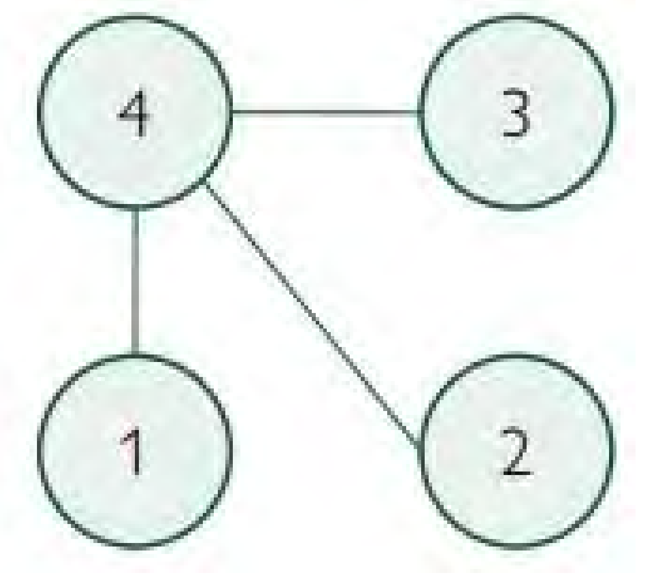
- Clique: Subset of nodes of a graph, such that all nodes are adjacent to each other
  - Triangles
  - Clique detection in graphs NP-Complete
  - Definition:
    - Usually too conservative/strict
    - Usually not found in sparse networks
    - May not reflect real social groups



Extension – K-Clique

[Wasserman & Faust 1994]

- K-Clique:
  - Maximal subgroup, where
  - largest geodesic distance between any pair of nodes is not greater than k



- 1-Clique is a clique
- 2-Clique: Subgraph, where all pairs of actors are connected with a path not longer than 2

Extension – Quasi-Clique

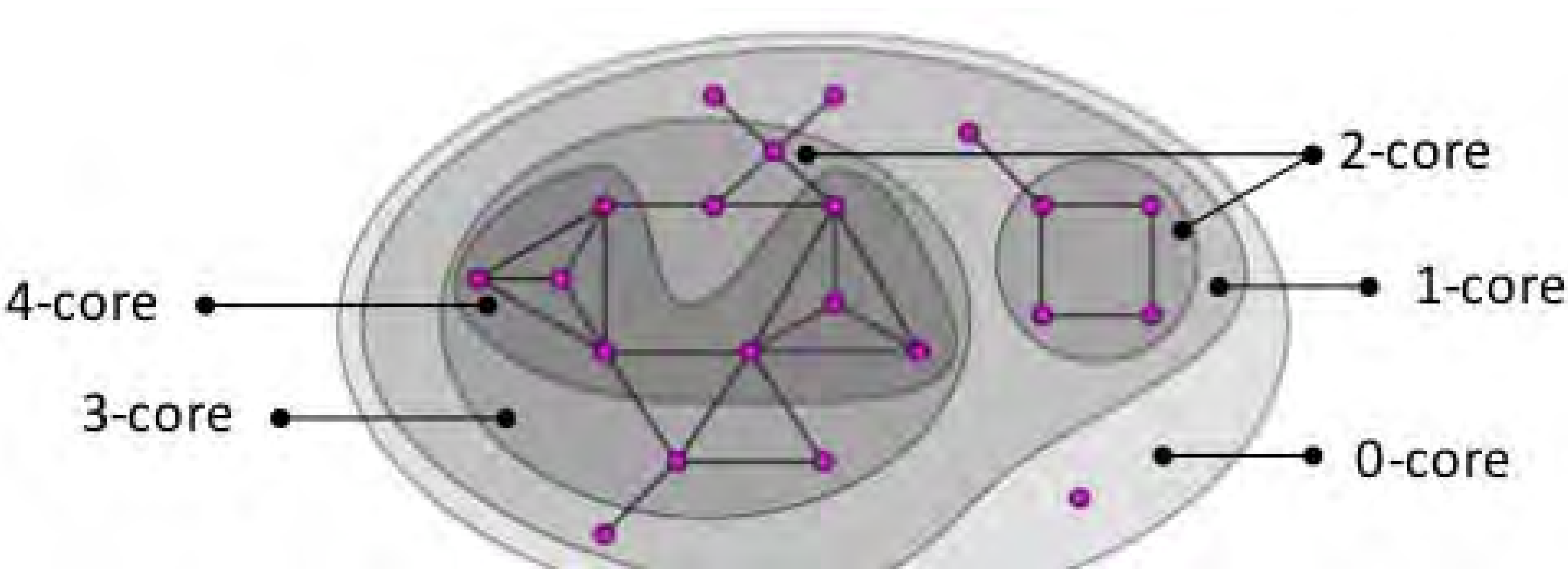
- Generalize clique to dense subgraph
- Different definitions (degree, density)
- Subset of nodes is quasi-clique, if
  - Nodal degree: every node in induced subgraph is adjacent to at least $\gamma(n-1)$ other nodes in the subgraph
  - Edge density: Number of edges in subgraph is at least $\lambda n(n-1)/2$

(with n : number of nodes in subgraph)

K-Core

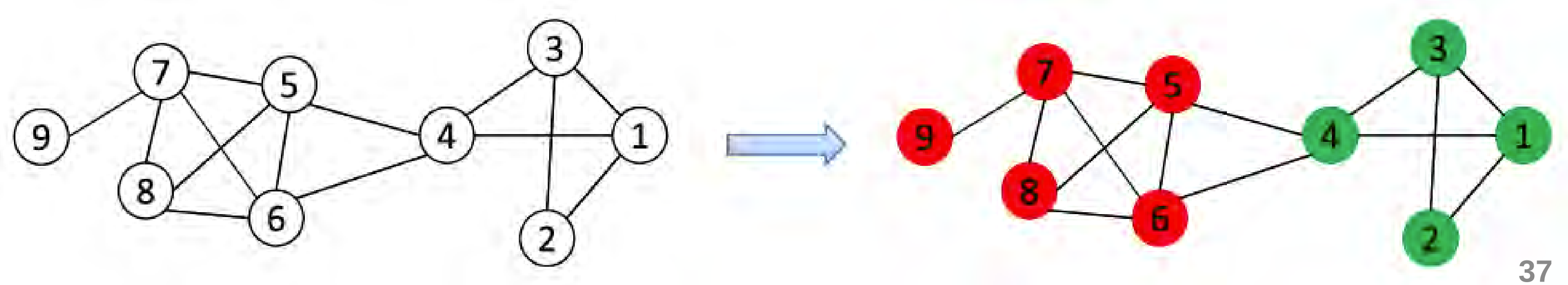
[Wasserman & Faust 1994]

- Maximal subgraph
- Each node has at least degree k
- Hierarchy of cores
  - Iteratively, eliminate lower-order cores
  - Until: Relatively dense subgroups remain



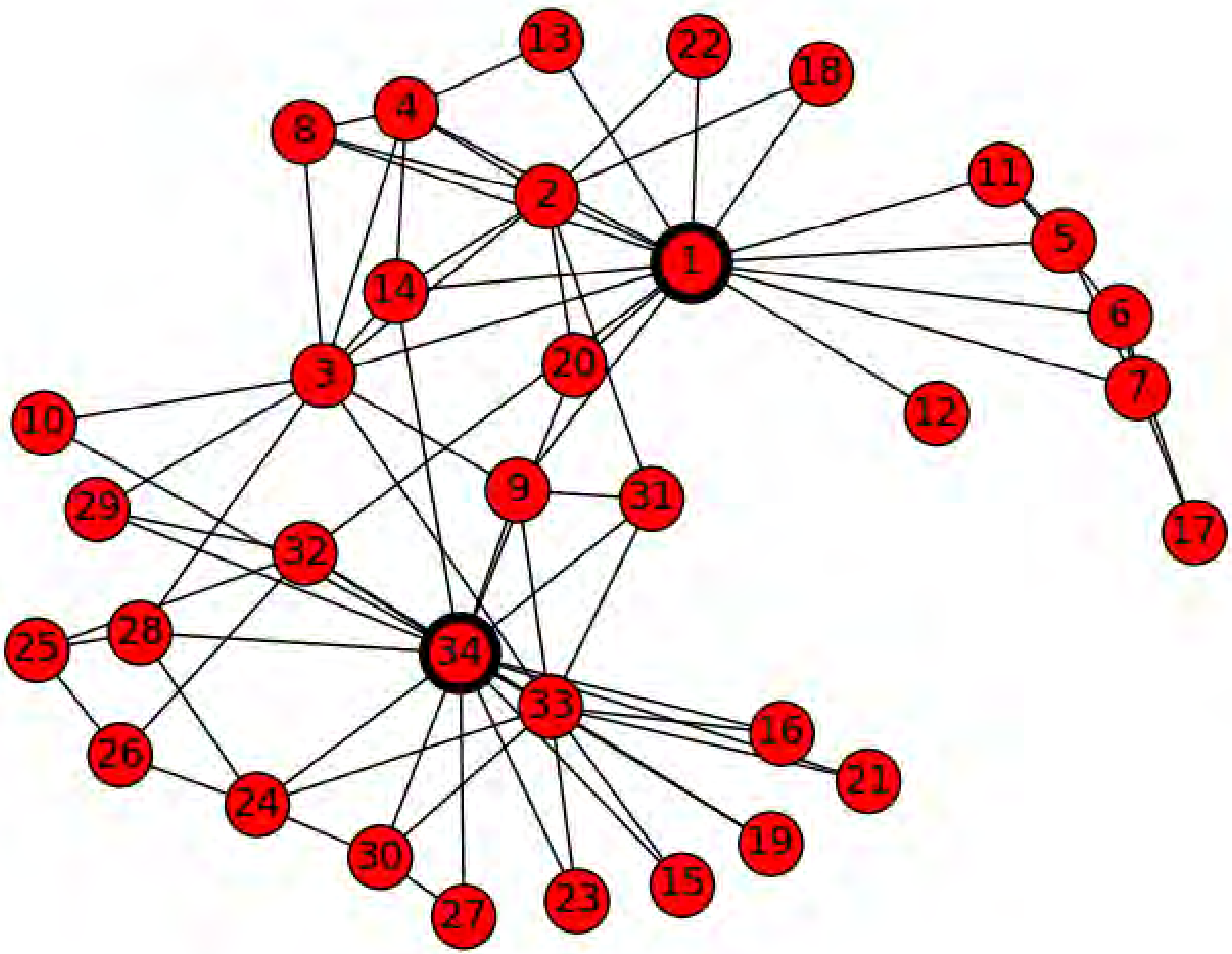
Communities

- Cohesive subgroups – structure within group
- Basic idea of communities
  - Tightly-knit groups
  - Consider both internal and external ties in network
  - In general:
    - High number of internal ties (high density within)
    - Low number of external ties (lower density between)



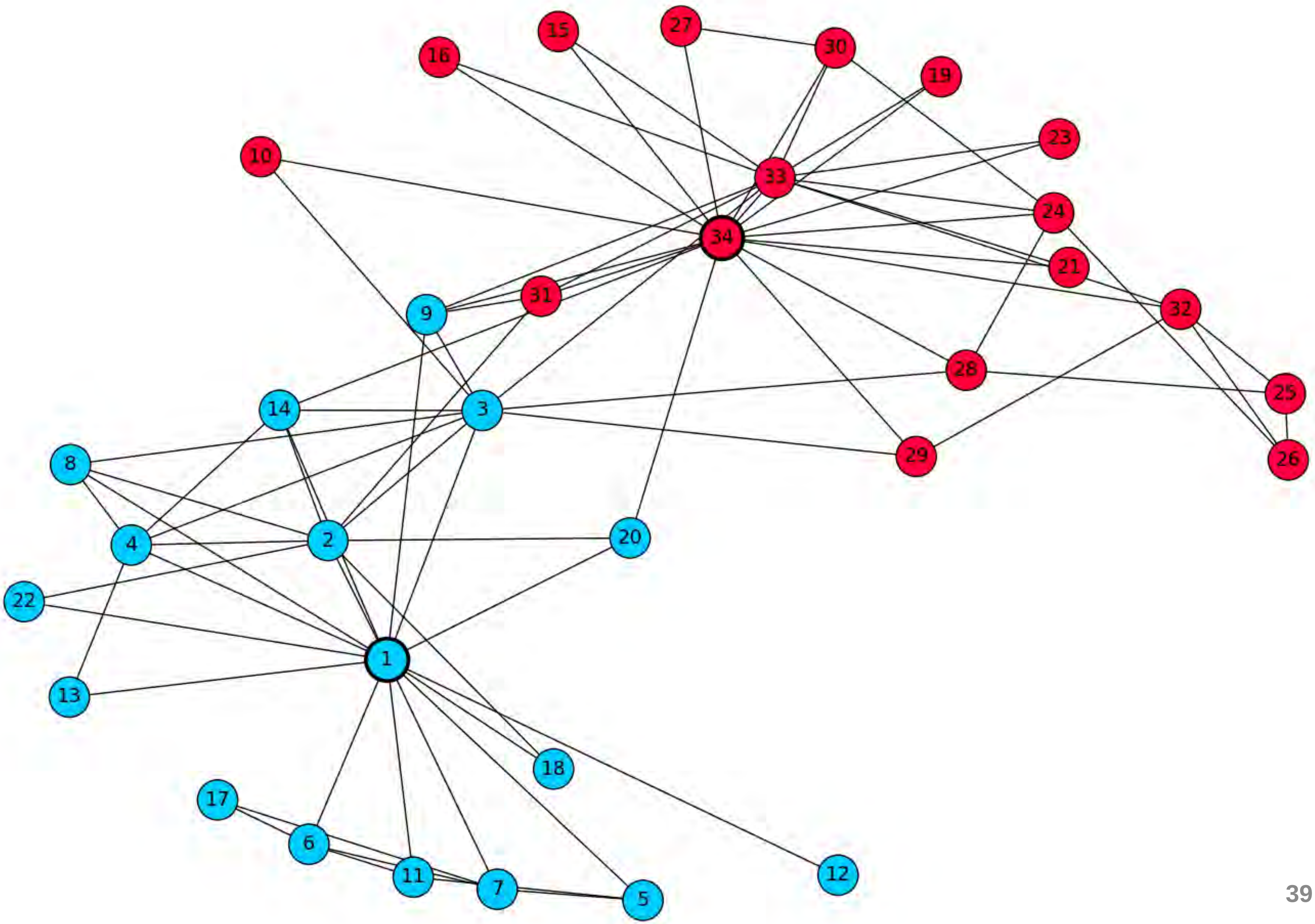
Zachary's Karate Club

[Zachary, 1977]



- Members of university karate club
- Conflict between club president (34) and karate instructor (1)
- Result: Split-up of the network according to friendship ties

Karate Club – 2 Factions



Finding Communities

- Given a network/graph, find "modules"
  - Single network [Newman 2002]
  - Multiplex networks [Bothorel 2015]
- Community structures [Fortunato 2010]
  - Graph Clustering → disjoint communities
  - Hierarchical organization [Lancichinetti 2009]
  - Overlapping communities [Xie et al. 2013]
- Questions:
  - What is "a community"?
  - What are "good" communities?
  - How do we evaluate these?

Community: Definition & Properties

- No universally accepted definition
- Informally:
  - Intuition: Densely connected group of nodes
  - Subset of nodes such that there are more edges inside the community than edges linking the nodes with the rest of the graph
- Intra Cluster Density
- Inter Cluster Density
- Connectedness

$$\delta_{int}(\mathcal{C}) = \frac{\#\text{ internal edges of }\mathcal{C}}{n_c(n_c-1)/2}$$

$$\delta_{ext}(\mathcal{C}) = \frac{\#\text{ inter-cluster edges of }\mathcal{C}}{n_c(n-n_c)}$$

Global View

- Communities can also be defined with respect to the whole graph
- Graph has community structure, if it is different from random graph
- Random graph: Not expected to have community structure
  - Here: Any two vertices have the same probability to be adjacent
  - Define null model; use it for investigating if we can observe community structure in a graph

Community Evaluation Measures

- Modularity [Newman 2006]

$$MOD(S) = \frac{1}{2m}\sum_{i,j}(A_{i,j} - \frac{d(i)d(j)}{2m})\delta(C_i, C_j)$$

Compares the number of edges within a community with the expected such number in a corresponding null model

$$MODL(C) = \frac{1}{2m}\sum_{i \in C, j \in C}(A_{i,j} - \frac{d(i)d(j)}{2m})$$

- Conductance [Kannan et al. 2004]

$$CON(C) = \frac{\overline{m}_C}{2m_C + \overline{m}_C}$$

Compares the number of edges within a community and the number of edges leaving the community

$$COIN(C) = 1 - CON(C) = \frac{2m_C}{\sum_{u \in C} d(u)}$$

Community Evaluation Measures

- Inverse Average Out-Degree Fraction (IAODF)

  [Leskovec et al. 2010]

$$\mathrm{IAODF}(C) := 1 - \frac{1}{n_C} \sum_{u \in C} \frac{\bar{\mathrm{d}}_C(u)}{\mathrm{d}(u)}$$

  compares the number of inter-edges to the number of all edges of a community, and averages this for the whole community by considering the fraction for each individual node

- Segregation Index (SIDX) [Freeman 1978]

$$\mathrm{SIDX}(C) = \frac{E(\bar{m}_C) - \bar{m}_C}{E(\bar{m}_C)} = 1 - \frac{\bar{m}_C n(n-1)}{2mn_C(n-n_C)}$$

  compares the number of expected interedges to the number of observed inter-edges, normalized by the expectation

Community Criteria [Tang & Liu 2010]

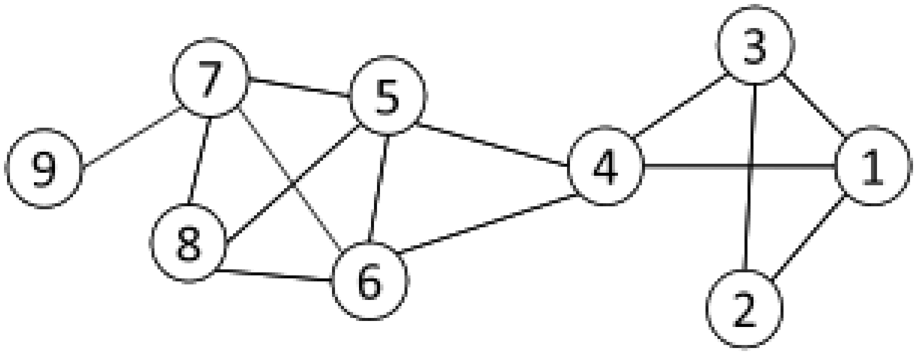
- Several possible community criteria
  - *Node*-Centric Community: Each node in a group satisfies certain properties, e.g., reachability, clique-based
  - *Group*-Centric Community: Consider the connections within a group as a whole. Group has to satisfy certain properties, e.g., minimal density, Quasi-clique …
  - *Network*-Centric Community: Partition the whole network into several disjoint sets, e.g., graph clustering, modularity maximization
  - *Hierarchy*-Centric Community: Construct a hierarchical structure of communities
  - *Descriptive* Community Detection: Identifies communities and description at the same time
    ➔ Especially for exploratory community detection

Clique Percolation Method (CPM)

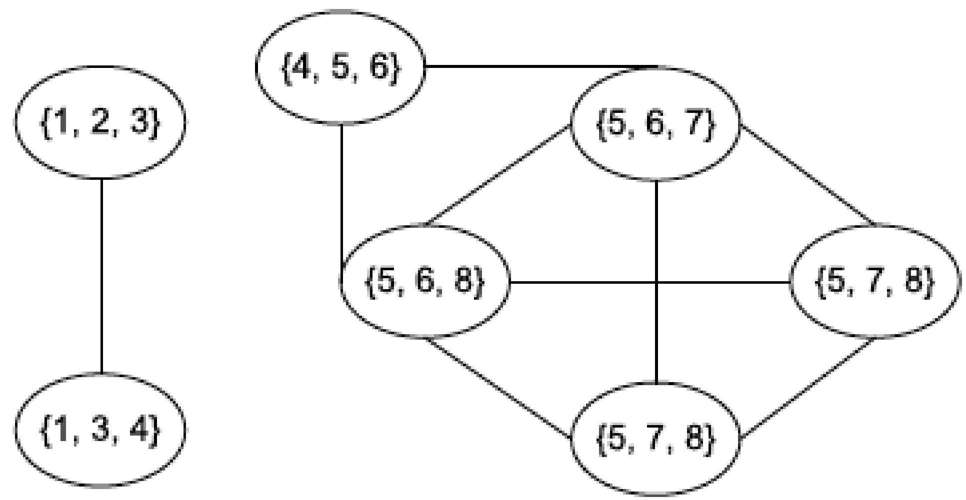
[Palla et al. 2005]

- Clique is a very strict definition, unstable
- Normally use cliques as a core or a seed to find larger communities
- CPM: Detect overlapping communities
  - **Input**
    - A parameter k, and a network
  - **Procedure**
    - Find out all cliques of size k in a given network
    - Construct a clique graph. Two cliques are adjacent if they share k-1 nodes
    - Each connected component in the clique graph forms a community

CPM Example



Cliques of size 3:
{1, 2, 3}, {1, 3, 4}, {4, 5, 6}, {5, 6, 7}, {5, 6, 8}, {5, 7, 8}, {6, 7, 8}



Communities:
{1, 2, 3, 4}
{4, 5, 6, 7, 8}

Network-Centric Community Detection

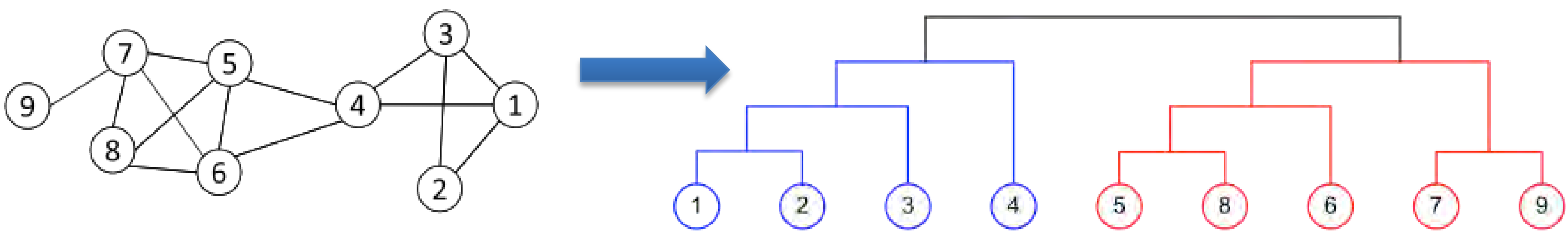
[Tang & Liu 2010]

- Network-centric criterion needs to consider the connections within a network globally
- Goal: partition nodes of a network into disjoint sets
- Approaches:
  - Clustering based on vertex similarity[Zhou et al. 2009]
  - Latent space models [Raftery et al. 2002]
  - Block model approximation [Karrer & Newman 2011]
  - Spectral clustering [Ma & Gao 2011]
  - Modularity maximization [Newman 2006]

Agglomerative Hierarchical Clustering

[Clauset et al. 2004]

- Initialize each node as a community
- Merge communities successively into larger communities following a certain criterion
  - E.g., based on modularity increase

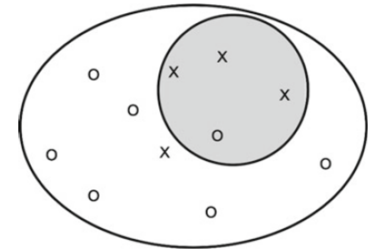


Divisive Hierarchical Clustering

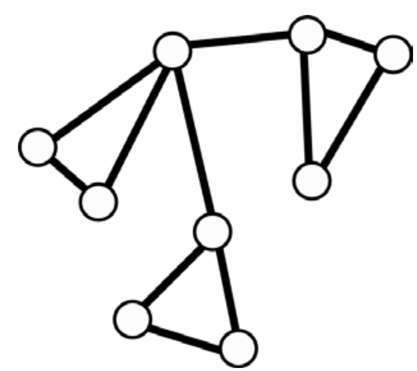
[Girvan & Newman 2002]

- Divisive clustering
  - Partition nodes into several sets
  - Each set is further divided into smaller ones
  - Network-centric partition can be applied for the partition
- One particular example: recursively remove the "weakest" tie
  - Find the edge with the least strength
  - Remove the edge and update the corresponding strength of each edge
- Recursively apply the above two steps until a network is discomposed into desired number of connected components.
- Each component forms a community

Agenda



- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook



Combining Structure and Attributes

- Data sources
  - Structural variables (ties, links)
  - Compositional variables
    - Actor attributes
    - Represented as attribute vectors
  - Edge attributes
    - Each edge has an assigned label
    - Multiplex graphs
      ➔ Multiple edges (labels) between nodes

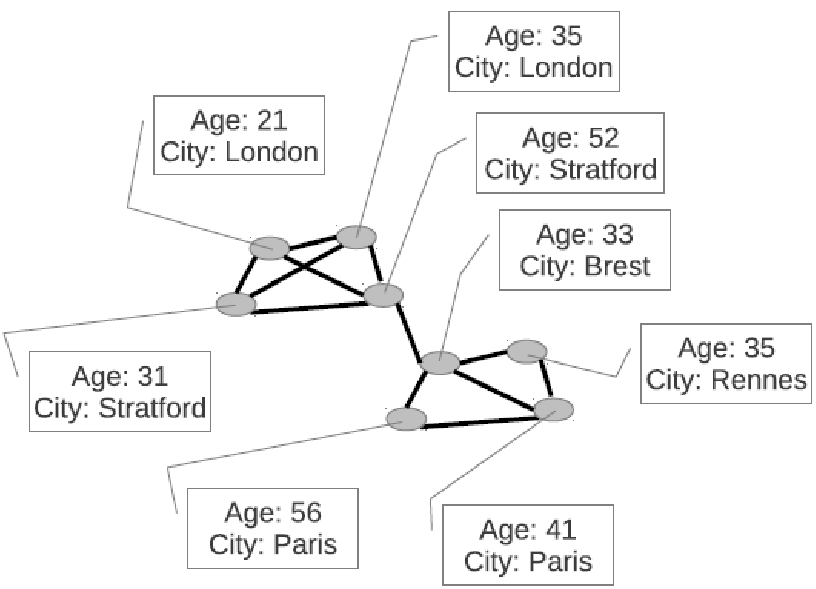
Communities/Edge-Attributed Graphs

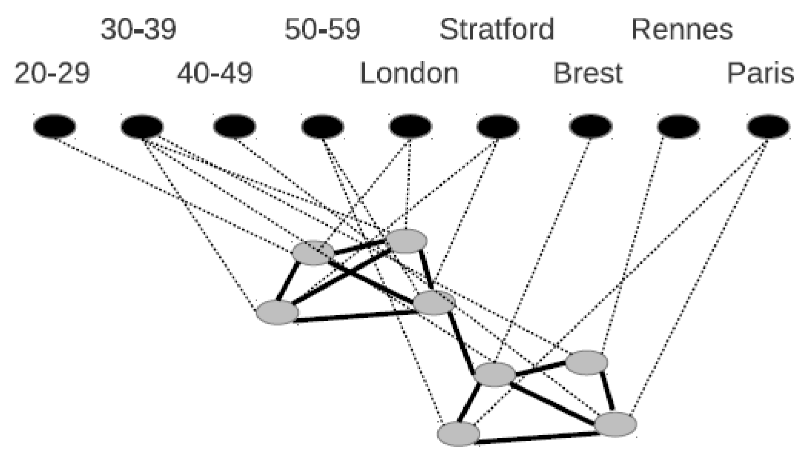
- Clustering edge-attributed graphs
  - Reduce/flatten to weighted graph [Bothorel et al. 2015]
    - Derive weights according to number of edges where nodes are directly connected [Berlingerio et al. 2011]
    - Standard graph clustering approaches can then be directly applied
  - Frequent-itemset based [Berlingerio et al. 2013]
  - Subspace-oriented [Boden et al. 2012]

Node-Attributed Graphs

- Non-uniform terminology
  - Social-attribute network
  - Attribute augmented graph
  - Feature-vector graph, vertex-labeled graph
  - Attributed graph
  - …
- Different representations

[Bothorel et al. 2015]





Community Detection – Attribute Extensions

- Utilize structural + attribute information
- Different roles of a description
  - Methods aiding community detection using attribute information
    - "Dense structures" - connectivity
    - But no "perfect" attribute homogeneity (purity)
  - Methods generating explicit descriptions, i.e., descriptive community patterns
    - "Dense structures" – connectivity
    - Concrete descriptions, e.g., conjunctive logical formula

Attributes for Aiding Community Detection

- Weight modification (edges) according to nodal attributes [Ge et al. 2008, Dang & Viennet 2012, Ruan et al. 2013, Zhou et al. 2009, Steinhaeuser & Chawla 2008]
  - Abstraction into similarities between nodes
    ➔ Edge weights
    ➔ Apply standard community detection algorithm,
  - Specifically, distance-based community detection methods
- Entropy-oriented methods [Zhu et al. 2011, Smith et al. 2014, Cruz et al. 2011]
- Model-based approaches [Xu et al. 2012, Yang et al. 2013, Akoglu et al. 2012]

Weight modification [Steinhaeuser & Chawla 2008]

- Use attribute-based distance measure

```
for each node i = 1...n do
    for each node j = 1...neighbors(i) do
        w(i, j) = 0
        for each node attribute a do
            if a is nominal and i.a = j.a then
                w(i, j) = w(i, j) + 1
            else if a is continuous then
                w(i, j) = w(i, j) + 1 − α|i.a − j.a|
            end if
        end for
    end for
end for
```

- Community detection: Group nodes according to threshold t, i.e., given $t \in (0, 1)$ place any pair of nodes whose edge weight exceeds the threshold into the same community
- Evaluate final partitioning using Modularity

Entropy Minimization

[Cruz et al. 2011]

- For a partition, optimize entropy using Monte-Carlo
- Integrate entropy step into Modularity optimization algorithm
  [Blondel et al. 2008]

Require: $\mathcal{C}$, imax, $PoV_{F_V^*}$

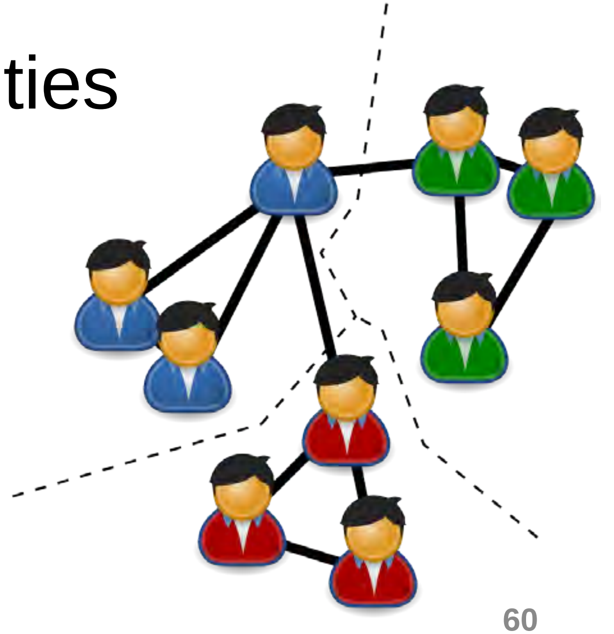
$H_0 \leftarrow \mathcal{H}_{\mathcal{C}}^0$
$i \leftarrow 0$
while $i < imax$ **and** More possible changes **do**
    $i \leftarrow i + 1$
    $A \leftarrow$random cluster from $\mathcal{C}$
    $x \leftarrow$random node : $x \in A$
    $A\,(x, -)$
    $B \leftarrow$random cluster from $\mathcal{C} \setminus \{\mathcal{D}_A \cup A\}$
    $B\,(x, +)$
    $H_i \leftarrow \mathcal{H}_{\mathcal{C}}^i$
    if $H_i \geq H_{i-1}$ **then**
        $B\,(x, -)$
        $A\,(x, +)$
    end if
end while
return $\mathcal{C}_{\mathcal{H}}$ {A new partition with a reduced entropy}

Model-based/MDL

- In general: Model edge & attribute values using mixtures of probability distributions
- Use MDL to select clusters w.r.t. attribute value similarity & connectivity similarity
  - Data compression of connectivity & attribute matrices (PICS algorithm) [Akoglu et al. 2013]
  - Lossless compression ➔ MDL cost-function
  - Resulting node groups
    - Homogeneous both in node & attribute matrix
    - Nodes - similar connectivity & high attribute coherence

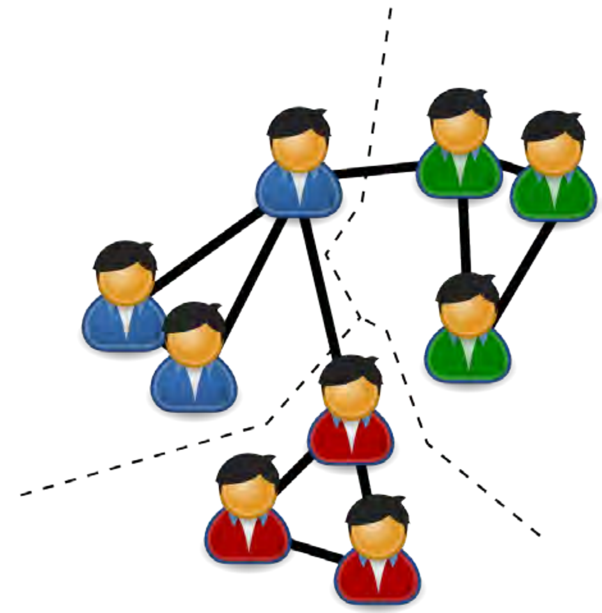
Descriptive Community Patterns

- Community mining scenario
  - Discover "densely connected groups of nodes"
  - Communities should have **explicit description**
  - Community (evaluation) space: network/graph
- Goal:
  - Often: Discover top-k communities
  - Maximize some community quality function



Examples: Community Patterns

- Social tagging system:
  - {work, flickr, delicious}
  - {business, production, sales}
  - {php, web, internet}, {innovation, business, forschung}
  - {work, flickr, delicious}, {library, android, emulation}, {php, web, internet}



Finding Explicit Descriptions

- Cluster transformed node-attribute similarity graph & extract pure clusters
- Mine frequent itemsets (binary attributes) & analyze communities [Adnan et al. 2009]
- Combine dense subgraph mining + subspace clustering [Moser et al. 2009,Günnemann et al. 2013]
- Apply correlated pattern mining [Silva et al. 2012]
- Interleave community detection & redescription mining [Pool et al. 2014]
- Adapt local exceptionality detection (using subgroup discovery) for communities [Atzmueller & Mitzlaff 2011, Atzmueller et al. 2015]

Subspace-Clustering & Dense Subgraphs

[Günnemann et al. 2011]

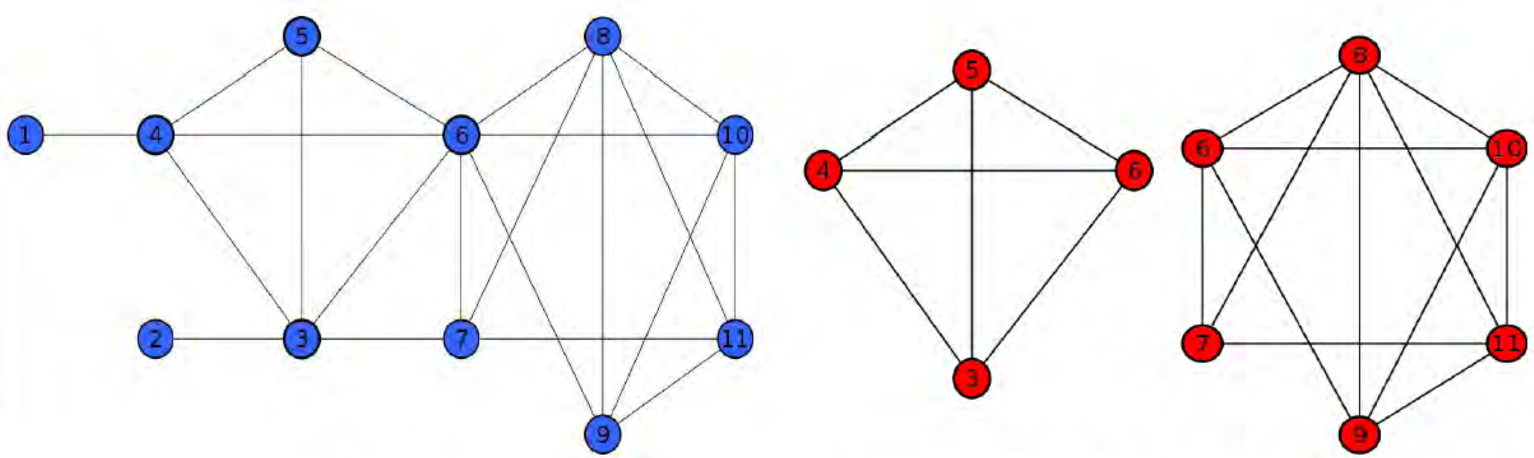
- Twofold cluster *O*: Combine subspace-clustering & dense subgraph mining (GAMer algorithm)
  - *O* fulfills subspace property (maximal distance threshold w.r.t. node attribute values in *O*) with minimal number of dimensions
  - *O* fulfills quasi-clique property, according to nodal-degree and threshold γ
  - Induced subgraph of *O* is connected, and fulfills minimal size threshold
- Quality function: Density · Size · #Dimensions
- Pruning using subspace & quasi-clique properties
- Includes Redundancy-optimization step (Overlapping communities)

Correlated Pattern Mining [Silva et al. 2011]

- Structural correlation pattern mining (SCPM)
  - Correlation between node attribute set and dense subgraph, induced by the attribute set
  - Quality measure: Comparison against null model
    - Size of the pattern
    - Cohesion of the pattern (density of quasi-clique)
  - Compare against expected structural correlation of attribute set (in random graph)

vertex	attributes
1	A, C
2	A
3	A, C, D
4	A,D
5	A, E
6	A, B, C
7	A, B, E
8	A, B
9	A, B
10	A, B, D
11	A, B

(a) Vertex attributes



(b) Graph (c) Dense subgraph (d) Dense subgraph

Algorithm 2 SCPM Algorithm

Require: $\mathcal{G}$, σ_{min}, γ_{min}, min_size, ϵ_{min}, δ_{min}, k
Ensure: $\mathcal{P}$

$\mathcal{P} \leftarrow \emptyset$
$\mathcal{T} \leftarrow \emptyset$
$\mathcal{I} \leftarrow$ *frequent attributes from* $\mathcal{G}$
for all $S \in \mathcal{I}$ **do**
  $\epsilon \leftarrow$ *structural correlation of* S
  if $\epsilon \geq \epsilon_{min}$ AND $\epsilon/\epsilon_{exp}(S) \geq \delta_{min}$ **then**
    $\mathcal{Q} \leftarrow$ *top-k patterns from* $\mathcal{G}(S)$
    for all $q \in \mathcal{Q}$ **do**
      $\mathcal{P} \leftarrow \mathcal{P} \cup (S, q)$
    end for
  end if
  if $\epsilon.\sigma(S) \geq \epsilon_{min}.\sigma_{min}$ AND $\epsilon.\sigma(S) \geq \delta_{min}.\epsilon_{exp}(\sigma_{min}).\sigma_{min}$ **then**
    $\mathcal{T} \leftarrow \mathcal{T} \cup S$
  end if
end for
$\mathcal{P} \leftarrow \mathcal{P} \cup$ *enumerate-patterns*$(\mathcal{T}, \mathcal{G}, \sigma_{min}, \gamma_{min}, min\_size, \epsilon_{min}, \delta_{min}, k)$

- Thresholds: min. support (size), structural correlation, expected structural correlation

Description-driven Community Detection

[Pool et al. 2014]

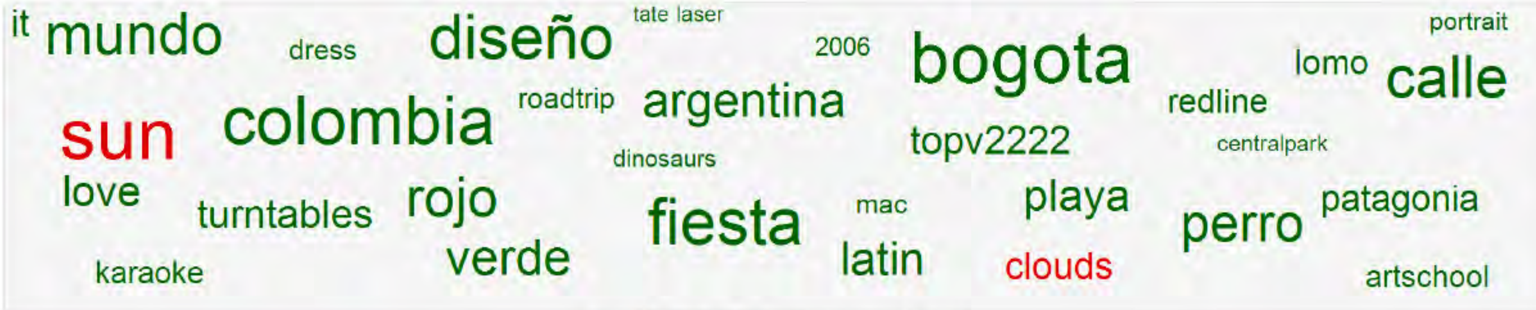
- Find communities with concise descriptions (e.g., given by tags)
- Focus: Overlapping, diverse, descriptive communities
- Language: Disjunctions of conjunctive expressions
- Two-stage approach
  - Greedy hill-climbing step: Generate candidates for communities
  - Redescription generation: Induce description for each community, and reshape if necessary
- Heuristic approach, due to large search space

ALGORITHM 1: DCM

Input: Attributed graph G, parameters k and η, and a set of candidate communities $\mathcal{C}$.
Output: An approximation of $\mathcal{Q}$, the top-k communities.

$\mathcal{Q} \leftarrow \emptyset$
for all $C \in \mathcal{C}$ **do**
  while C changes **do**
    $C \leftarrow$ MAXIMIZE_COMMUNITY_SCORE(C)
    $C \leftarrow$ FIND_CONCISE_QUERY(C)
  end while
  $\mathcal{Q} \leftarrow \mathcal{Q} \cup Q(C)$
end for
$\mathcal{Q} \leftarrow$ SELECT_DIVERSE_TOP_K($\mathcal{Q}, k, \eta$)
return $\mathcal{Q}$

- Starts with candidate communities
  - Domain knowledge
  - Partial communities
  - Start with single vertices (later being extended using hill-climbing approach)
- ReMine algorithm for deriving patterns for communities [Zimmermann et al. 2010]



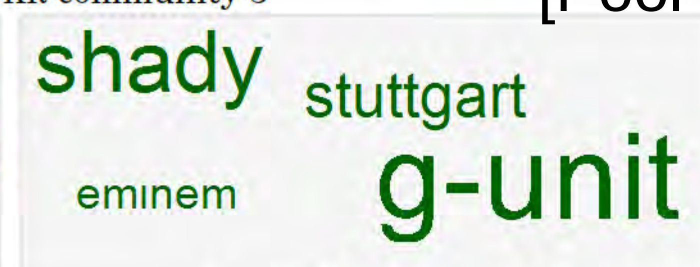
a. FLICKR community 1

b. FLICKR community 2

c. FLICKR community 3

[Pool et al. 2013]

d. LASTFM community 1

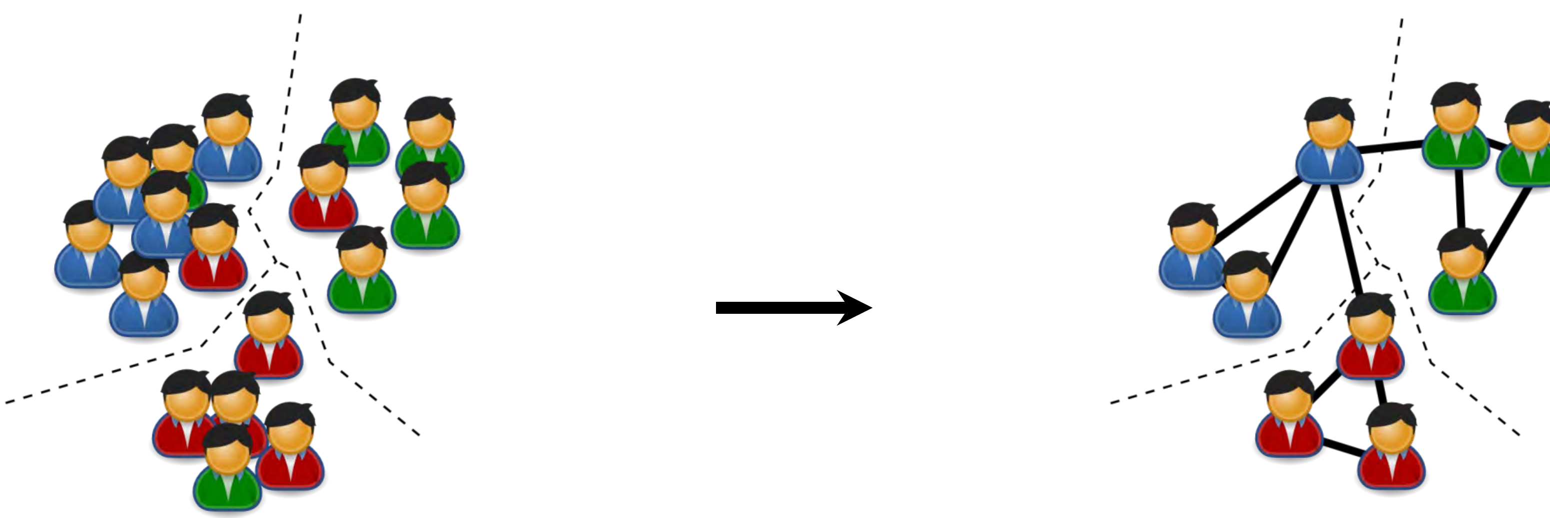


e. LASTFM community 2

Description-Oriented Community Detection

[Atzmueller et al., Information Science, 2016]

- Basic Idea: Pattern Mining for Community Characterization
  - Mine patterns in description space (tags/topics)
    ➔ Subgroups of users **described** by tags/topics
  - Optimize quality measure in community space
    ➔ Network/graph of users
  - Improve understandability of communities (explanation)



Direct Descriptive Community Mining

- Goal: Identification/description of communities with a high quality (exceptional model mining)
  - **Input:** Network/Graph + node properties (e.g., tags)
  - **Output:** k-best community patterns
- Description language: conjunctive expressions
- COMODO algorithm: Top-k pattern mining, based on SD-Map* algorithm for subgroup discovery
  - Discover k-best patterns
  - Search space: Conjunctions/tags
  - Apply standard community quality functions, e.g., Modularity [Newman 2004]

$$MOD(S) = \frac{1}{2m} \sum_{i,j} (A_{i,j} - \frac{d(i)d(j)}{2m}) \delta(C_i, C_j)$$

Community Detection on Attributed Graphs

- Goal: Mine patterns describing such groups

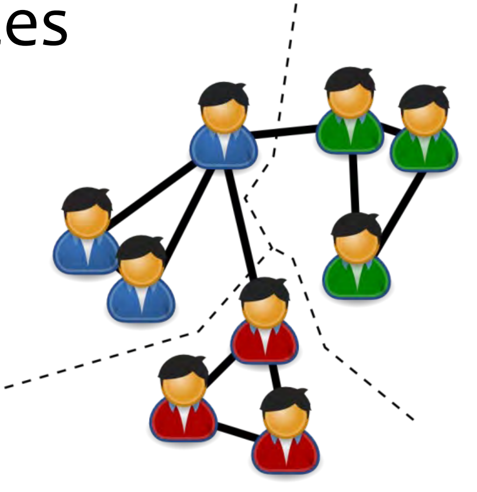
Size	Community description	Size	Community description
519	80s	32	psychedelic AND minimal
240	gregorian_chant AND 80s	16	psychedelic AND 80s
215	girl_groups AND 80s	10	psychedelic AND brit_rock AND classic_rock
171	atmospheric	10	death_rock AND minimal AND 80s
122	synth_pop	10	death_rock AND 80s AND doom_metal

  - Merge networks + descriptive features, e.g., characteristics of users
  - Target both
    - Community structure (some evaluation function) &
    - Community description (logical formula, e.g., conjunction of features, see above)

Transformation & Mining (I)

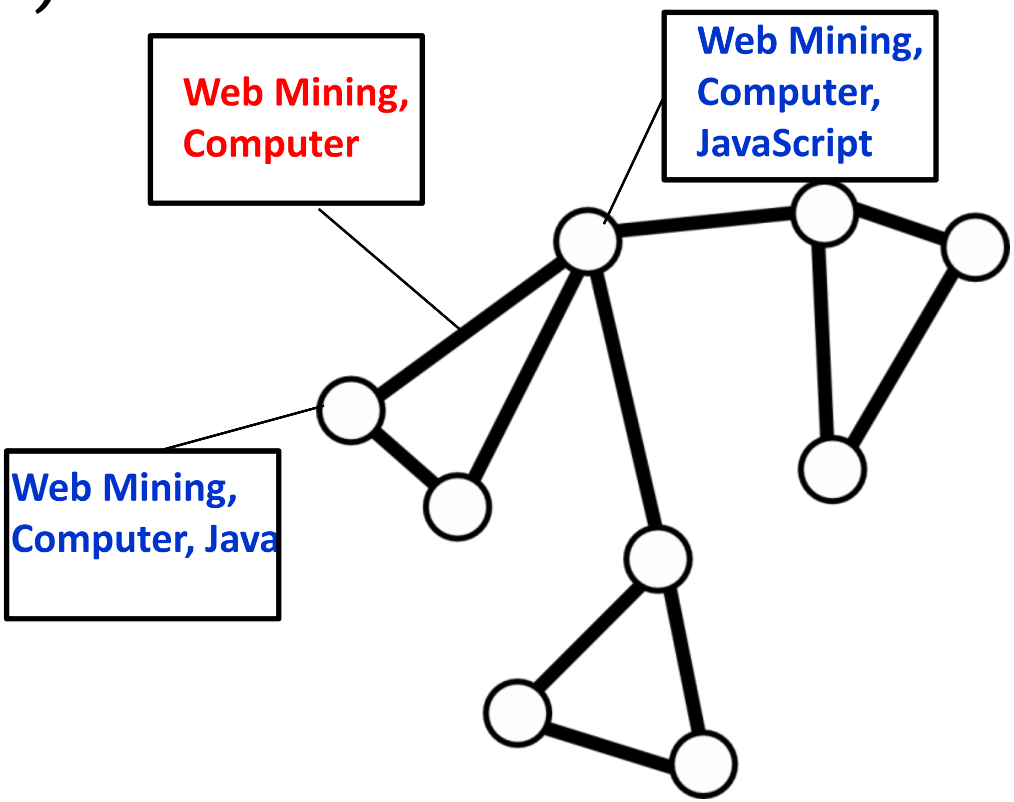
- Sources:
  - Database DB: Users described by attributes (e.g. used topics)
  - Graph G: Links between users (e.g. friend graph)
- Goal:
  - Discover k best communities as subgroups of DB
  - Maximizing community evaluation function on G
- Need to merge both data sources

User 1: {work, flickr, delicious}
User 2: {business, production, sales}
User 3: {php, web, internet},
{innovation, business, forschung}
User 4: {work, flickr, delicious},
{library, android, emulation},
{php, web, internet}
…



Transformation & Mining (II)

- Dataset of edges connecting two nodes
  - Described by intersection of labels of the two nodes
  - Additionally: Store nodes, and respective degrees
- Apply top-k method w/ optimistic-estimate pruning (COMODO)



Algorithm 1 COMODO

procedure COMODO-Mine (cf. [17] for an extended description)

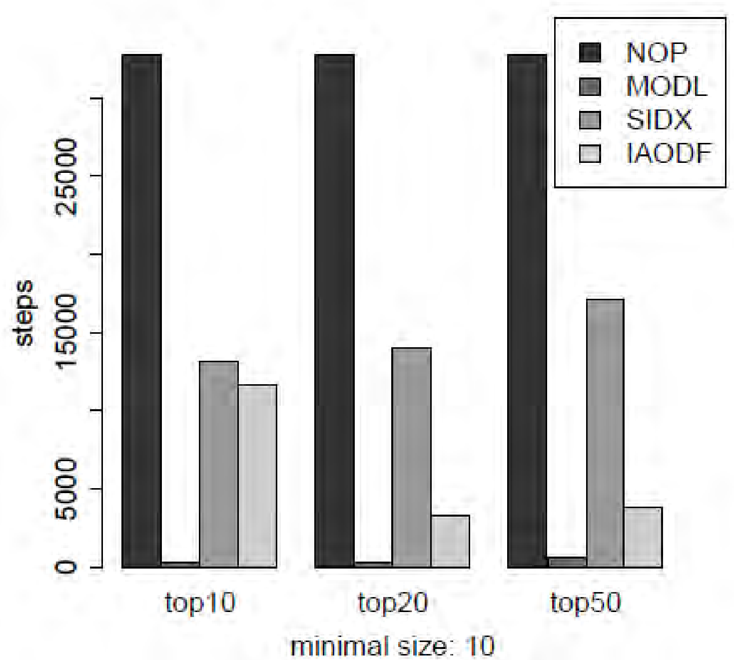
Input: Current community pattern tree CPT, pattern $\hat{p}$, priority queue *top-k*, int k (max. number of patterns), int $maxLength$ (max. length of a pattern), int τ_n (min. community size)

```
COM = new dictionary: basicpattern → pattern
minQ = minQuality(top-k)
for all b in CPT.getBasicPatterns do
    p = createRefinement(p̂, b)
    COM[b] = p
    if size(p, CPT) ≥ τ_n then
        if quality(p, F) ≥ minQ then
            addToQueue(top-k, p)
            minQ = minQuality(top-k)
if length(p̂) + 1 < maxLength then
    refinements = sortBasicPatternsByOptimisticEstimateDescending(COM)
    for all b in refinements do
        if optimisticEstimate(COM[b]) ≥ minQ then
            CCPT = getConditionalCPT(b, CPT, minQ)
            Call COMODO-Mine(CCPT, COM[b], top-k)
```

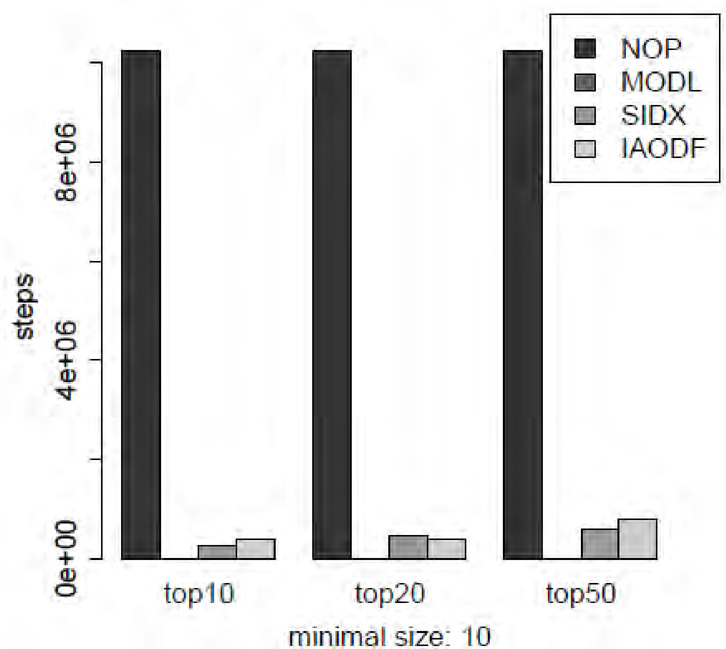
- Algorithm utilizes special tree-structure & optimistic estimates for efficient processing

Optimistic Estimates

- Problem: Exponential Search Space
- Optimistic Estimate: Upper bound for the quality of a pattern and all its specializations
  ➔Top-K Pruning



Last.fm friend graph



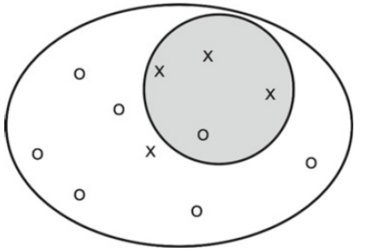
Delicious friend graph

Optimistic Estimate Pruning

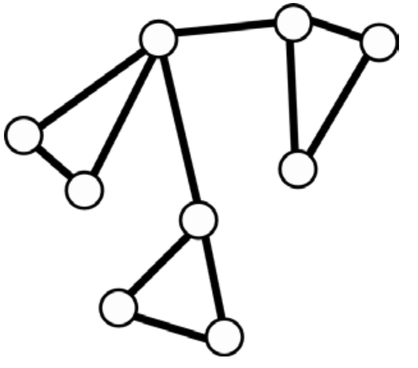
[Atzmueller et al. 2015]

Friend Graph, LDA-100						
Method	Top k	$\tau_n = 5$	$\tau_n = 10$	$\tau_n = 20$	$\tau_n = 50$	$\tau_n = 70$
MODL	5	619	619	619	87	15
	10	1,654	1,654	1,654	87	15
	20	4,104	4,104	4,104	87	15
	50	48,460	30,287	47,945	87	15
dMODL	5	420	420	420	87	15
	10	1,393	1,393	1,393	87	15
	20	3,479	3,688	3,676	87	15
	50	34,620	51,041	50,236	87	15
COIN	5	79,219,778	60,421,374	1,215,426	86	15
	10	79,219,778	61,722,695	1,373,936	87	15
	20	79,219,778	62,340,121	1,463,930	87	15
	50	79,219,778	63,052,402	1,555,210	87	15
No Pruning		79,375,495	69,971,088	3,205,222	87	15

Agenda



- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook



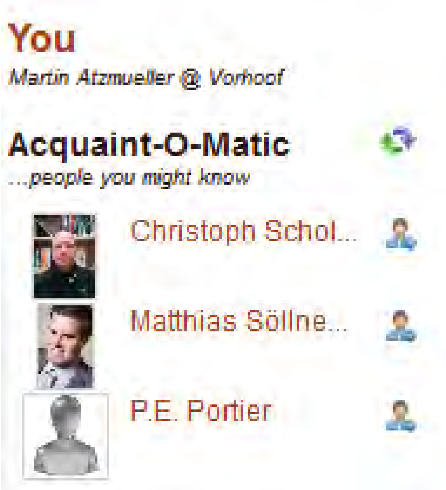
Descriptive Community Detection

- Example: Patterns from last.fm
  - Recommendation
  - Browsing
  - …

Size	Community description
519	80s
240	gregorian_chant AND 80s
215	girl_groups AND 80s
171	atmospheric
122	synth_pop
32	psychedelic AND minimal
16	psychedelic AND 80s
10	psychedelic AND brit_rock AND classic_rock
10	death_rock AND minimal AND 80s
10	death_rock AND 80s AND doom_metal

Conferator

- Interest profiles – Recommending conference participants
  - ➔ BibSonomy: User profiles
    - java AND android AND nfc
  - ➔ Conferator: Acquaint-O-Matic
    - recommender & analysis
    - mining & ubiquitous & social



Further Examples

- Behavioral social targeting
  - Apply domain knowledge
  - Use (explicit) descriptions
- Recommendations
  - Popular items in community
  - Deal with cold-start problems
- Exploratory analytics
  - First insights into data
  - Characterization of exceptional subgroups

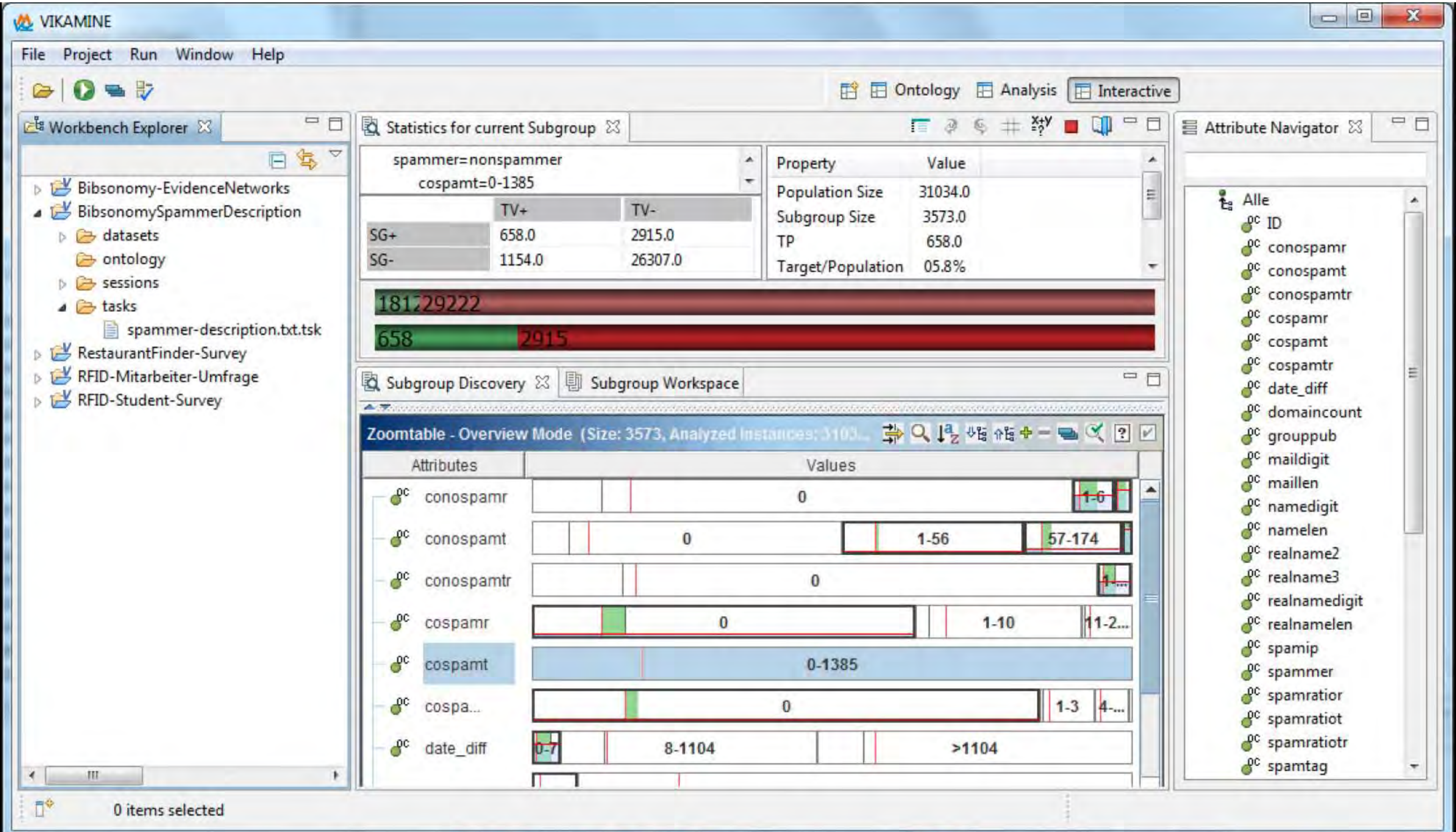
Tool: VIKAMINE

- **V**isual, **I**nteractive and **K**nowledge-intensive **A**nalysis and semantic **MIN**ing Environment
  - Data mining
  - Visual analytics
  - Knowledge refinement
  - Semantic knowledge capture
- Option: Include background knowledge, semantic annotation, ontologies
- http://www.vikamine.org
  (R-Package: rsubgroup.org)

VIKAMINE Features

- Efficient automatic discovery algorithms
  - Subgroup discovery & analytics
  - Community detection
- Seamless integration of visualization methods
- Effective visualizations for ad-hoc analysis
- Ad-hoc formalization, utilization, and extension of background knowledge
- Plugin for description-oriented community detection
- Works also on big data (Map/Reduce)

Workbench



www.vikamine.org

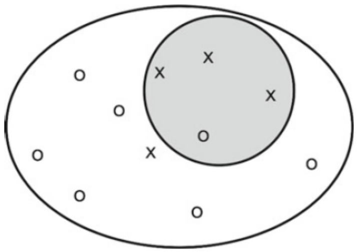
Community Detection Software

- igraph (R):
  - Different community detection methods (mostly methods for detecting disjoint communities)
    - Modularity maximization, Walktrap, Label propagation, INFOMAP, …
- Linkcomm (R): detection of link communities (potentially overlapping)
- CPM
  - CFinder: http://www.cfinder.org/
  - Fast clique percolation (cp5) : https://github.com/aaronmcdaid/MaximalCliques

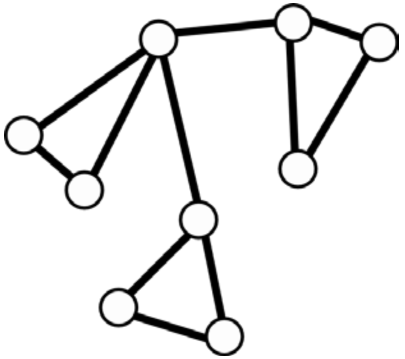
Community Detection Software

- SNAP (**S**tanford **N**etwork **A**nalysis **P**latform): https://snap.stanford.edu/snap/description.html
- Overlapping communities:
  - COPRA: http://www.cs.bris.ac.uk/~steve/networks/copra/
  - MOSES: http://sites.google.com/site/aaronmcdaid/moses
- Description-oriented methods/attributed graphs
  - COMODO: www.vikamine.org
  - DCM: http://www.patternsthatmatter.org/software.php#dcm
  - GAMER: http://dme.rwth-aachen.de/de/gamer
- Bipartite networks: http://danlarremore.com/bipartiteSBM/

Agenda



- Motivation
- Basics: Graphs & Attributes
- Subgroup Discovery & Analytics
- Cohesive Subgroups & Communities
- Community Detection on Attributed Graphs
- Applications & Tools
- Summary & Outlook



Summary

- Subgroup discovery & community detection enable the identification of subgroups at different levels & dimensions
  - Compositional
  - Structural + compositional
  - Providing explicit descriptions
- Both can be combined for obtaining descriptive community patterns according to standard community quality functions
- Efficient tools for detection & analysis

Outlook

- Challenges using ubiquitous & social data
  - Heterogeneous data & complex networks
  - Integration of multiplex networks & temporal information
  - Support for integration & analysis
  - Necessary: Efficient methods and tools for the mining of such data
- Extensions: Effective exploratory methods for analytics. Integrated assessment, mining & inspection

Community Detection: From Plain to Attributed Complex Networks

Martin Atzmueller

University of Kassel, Research Center for Information System Design
Ubiquitous Data Mining Team, Chair for Knowledge and Data Engineering

Web Science 2016, Hannover – 2016-05-22

References (cont.)

- [Atzmueller et al. 2005] M. Atzmueller, F. Puppe, and H.-P. Buscher (2005) Profiling Examiners using Intelligent Subgroup Mining. In Proc. 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005), pp. 46-51, Aberdeen, Scotland
- [Atzmueller & Puppe 2008] M. Atzmueller and F. Puppe (2008) A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence, 28(3):210-221
- [Atzmueller & Hilgenberg 2013] M. Atzmueller and K. Hilgenberg (2013) Towards Capturing Social Interactions with SDCF: An Extensible Framework for Mobile Sensing and Ubiquitous Data Collection. In Proc. 4th International Workshop on Modeling Social Media (MSM 2013), Hypertext 2013, New York, NY, US. ACM Press.
- [Atzmueller & Lemmerich 2012] M. Atzmueller and F. Lemmerich (2012) VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In Proc. ECML/PKDD 2012: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Heidelberg, Germany. Springer Verlag.
- [Atzmueller & Puppe 2005] M. Atzmueller and F. Puppe (2005) Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science, 11(11):1752-1765, 2005.
- [Atzmueller & Lemmerich 2009] M. Atzmueller, F. Lemmerich (2009) Fast Subgroup Discovery for Continuous Target Concepts. Proc. International Symposium on Methodologies for Intelligent Systems, Vol. 5722 of LNCS, Springer, Berlin, pp. 1–15.
- [Atzmueller et al. 2012] M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme (2012) Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In Modeling and Mining Ubiquitous Social Media, volume 7472 of LNAI. Springer Verlag, Heidelberg, Germany

References

- [Adnan et al. 2009] M. Adnan, R. Alhajj, J. Rokne (2009) Identifying Social Communities by Frequent Pattern Mining. Proc. 13th Intl. Conf. Information Visualisation, IEEE Computer Society, Washington, DC, USA, pp. 413–418.
- [Akoglu et al. 2012] L. Akoglu, H. Tong, B. Meeder, and C. Faloutsos (2012) Pics: Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs. Proc. SDM, SIAM, pp. 439–450. Omnipress
- [Atzmueller 2015] Atzmueller, M (2015) Subgroup Discovery – Advanced Review. WIREs: Data Mining and Knowledge Discovery, 5(1):35–49
- [Atzmueller 2007] M. Atzmueller (2007) Knowledge-Intensive Subgroup Mining – Techniques for Automatic and Interactive Discovery, Vol. 307 of Dissertations in Artificial Intelligence-Infix (Diski), IOS Press
- [Atzmueller et al. 2004] M. Atzmueller, F. Puppe, H.-P. Buscher (2004) Towards Knowledge-Intensive Subgroup Discovery, Proc. LWA 2004, pp. 117–123.
- [Atzmueller & Puppe 2006] M. Atzmueller and F. Puppe (2006) SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006), pp. 6-17, Heidelberg, Germany. Springer Verlag
- [Atzmueller et al. 2005] M. Atzmueller, J. Baumeister, A. Hemsing, E.-J. Richter, and F. Puppe (2005) Subgroup Mining for Interactive Knowledge Refinement. In Proc. 10th Conference on Artificial Intelligence in Medicine AIME 05), LNAI 3581, pp. 453-462, Heidelberg, Germany, Springer Verlag.

References (cont.)

- [Atzmueller et al. 2005] M. Atzmueller, F. Puppe, and H.-P. Buscher (2005) Profiling Examiners using Intelligent Subgroup Mining. In Proc. 10th International Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005), pp. 46-51, Aberdeen, Scotland

- [Atzmueller & Puppe 2008] M. Atzmueller and F. Puppe (2008) A Case-Based Approach for Characterization and Analysis of Subgroup Patterns. Journal of Applied Intelligence, 28(3):210-221

- [Atzmueller & Hilgenberg 2013] M. Atzmueller and K. Hilgenberg (2013) Towards Capturing Social Interactions with SDCF: An Extensible Framework for Mobile Sensing and Ubiquitous Data Collection. In Proc. 4th International Workshop on Modeling Social Media (MSM 2013), Hypertext 2013, New York, NY, US. ACM Press.

- [Atzmueller & Lemmerich 2012] M. Atzmueller and F. Lemmerich (2012) VIKAMINE - Open-Source Subgroup Discovery, Pattern Mining, and Analytics. In Proc. ECML/PKDD 2012: European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, Heidelberg, Germany. Springer Verlag.

- [Atzmueller & Puppe 2005] M. Atzmueller and F. Puppe (2005) Semi-Automatic Visual Subgroup Mining using VIKAMINE. Journal of Universal Computer Science, 11(11):1752-1765, 2005.

- [Atzmueller & Lemmerich 2009] M. Atzmueller, F. Lemmerich (2009) Fast Subgroup Discovery for Continuous Target Concepts. Proc. International Symposium on Methodologies for Intelligent Systems, Vol. 5722 of LNCS, Springer, Berlin, pp. 1–15.

- [Atzmueller et al. 2012] M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme (2012) Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In Modeling and Mining Ubiquitous Social Media, volume 7472 of LNAI. Springer Verlag, Heidelberg, Germany

References (cont.)

- [Atzmueller & Lemmerich 2013] M. Atzmueller and F. Lemmerich (2013) Exploratory Pattern Mining on Social Media using Geo-References and Social Tagging Information. IJWS, 2(1/2)
- [Atzmueller & Mitzlaff 2011] M. Atzmueller and F. Mitzlaff (2011) Efficient Descriptive Community Mining. Proc. 24th International FLAIRS Conference, pages 459-464, Palo Alto, CA, USA. AAAI Press.
- [Atzmueller et al. 2015] M. Atzmueller, S. Doerfel, and F. Mitzlaff (2015) Description-Oriented Community Detection using Exhaustive Subgroup Discovery. Information Sciences. http://dx.doi.org/10.1016/j.ins.2015.05.008.
- [Atzmueller et al. 2009] M. Atzmueller, F. Lemmerich, B. Krause, and A. Hotho (2009) Who are the Spammers? Understandable Local Patterns for Concept Description. In Proc. 7th Conference on Computer Methods and Systems, Krakow, Poland. Oprogramowanie Nauko-Techniczne.
- [Berlingerio et al. 2013] M. Berlingerio, F. Pinelli, and F. Calabrese (2013) ABACUS: Apriori-BAsed Community discovery in mUltidimensional networkS. Data Mining and Knowledge Discovery, Springer, 27(3).
- [Boden et al. 2012] B. Boden, S. Günnemann, H. Hoffmann, and T. Seidl (2012) Mining Coherent Subgraphs in Multi-Layer Graphs with Edge Labels. Proc. 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press
- [Bothorel et al. 2015] C. Bothorel, J. D. Cruz, M. Magnani, B. Micenkova (2015) Clustering Attributed Graphs: Models, Measures and Methods. arXiv:1501.01676

References (cont.)

- [Bringmann et al. 2011] B. Bringmann, S. Nijssen, and A. Zimmermann (2011) Pattern-based Classification: A Unifying Perspective. arXiv:1111.6191

- [Clauset et al. 2004] A. Clauset, M. E. J. Newman, C. Moore (2004) Finding Community Structure in Very Large Networks. arXiv:cond-mat/0408187

- [Cruz et al. 2011] J. D. Cruz, C. Bothorel, F. and Poulet (2011) Entropy Based Community Detection in Augmented Social Networks. Computational Aspects of Social Networks, pp. 163-168

- [Dang & Viennet 2012] T. A. Dang and E. Viennet (2012) Community Detection Based on Structural and Attribute Similarities. Proc. International Conference on Digital Society (ICDS), pp. 7-14

- [Duivestein et al. 2015] W. Duivesteijn, A.J. Feelders, and A. Knobbe (2015) Exceptional Model Mining - Supervised Descriptive Local Pattern Mining with Complex Target Concepts. Data Mining and Knowledge Discovery

- [Fortunato 2010] S. Fortunato (2010) Community Detection in Graphs, Physics Reports 486 (3-5)

- [Freeman 1978] L. Freeman (1978) Segregation In Social Networks, Sociological Methods & Research 6 (4)

References (cont.)

- [Ge et al. 2008] R. Ge, M. Ester, B. J. Gao, Z. Hu, B. Bhattacharya, and B. Ben-Moshe (2008) Joint Cluster Analysis of Attribute Data and Relationship Data: The Connected k-Center Problem, Algorithms and Applications. Acm Trans. Knowl. Discov. Data, 2(2)
- [Girvan & Newman 2002] M. Girvan, M. E. J. Newman (2002) Community Structure in Social and Biological Networks, PNAS 99 (12)
- [Günnemann et al. 2013] S. Günnemann, I. Färber, B. Boden, T. Seidl (2013) GAMer: A Synthesis of Subspace Clustering and Dense Subgraph Mining. Knowledge and Information Systems (KAIS), Springer
- [Kannan et al. 2004] R. Kannan, S. Vempala, A. Vetta (2004) On Clustering: Good, Bad and Spectral. Journal of the ACM, 51(3)
- [Kloesgen 1996] Klösgen, W. (1996) Explora: A Multipattern and Multistrategy Discovery Assistant. In Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., and Uthurusamy, R., editors, Advances in Knowledge Discovery and Data Mining, pp. 249–271. AAAI Press.
- [Lancichinetti 2009] A. Lancichinetti, S. Fortunato (2009) Community Detection Algorithms: A Comparative Analysis. arXiv:0908.1062
- [Lazarsfield & Merton 1954] P. F. Lazarsfeld, R. K. Merton (1954) Friendship as a Social Process: A Substantive and Methodological Analysis. Freedom and Control in Modern Society, 18(1), 18-66

References (cont.)

- [Leman et al. 2008] D. Leman, A. Feelders, and A. Knobbe (2008). Exceptional Model Mining. In Proc. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, volume 5212 of Lecture Notes in Computer Science, pages 1–16. Springer.
- [Lemmerich et al. 2012] F. Lemmerich, M. Becker, and M. Atzmueller (2012) Generic Pattern Trees for Exhaustive Exceptional Model Mining. In Proc. ECML/PKDD, Heidelberg, Germany. Springer
- [Leskovec et al. 2010] J. Leskovec, K. J. Lang, and M. Mahoney (2010) Empirical Comparison of Algorithms for Network Community Detection. Proc. 19th International Conference on World Wide Web, pp. 631-640. ACM
- [McPherson et al. 2011] M. McPherson, L. Smith-Lovin, and J. M. Cook (2001) Birds of a Feather: Homophily in Social Networks. Annual Review of Sociology, 415-444
- [Mitzlaff et al. 2011] F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme (2011) Community Assessment using Evidence Networks. In Analysis of Social Media and Ubiquitous Data, volume 6904 of LNAI
- [Mitzlaff et al. 2013] F. Mitzlaff, M. Atzmueller, D. Benz, A. Hotho, and G. Stumme (2013) User-Relatedness and Community Structure in Social Interaction Networks. CoRR/abs, 1309.3888
- [Moser et al. 2009] F. Moser, R. Colak, A. Rafiey, and M. Ester (2009) Mining Cohesive Patterns from Graphs with Feature Vectors. Proc. SDM (Vol. 9), pp. 593-604.

References (cont.)

- [Newman 2004] M. E. Newman (2004). Detecting community structure in networks. The European Physical Journal B-Condensed Matter and Complex Systems, 38(2), 321-330.
- [Newman 2006] M. E. Newman 2006) Modularity and Community Structure in Networks. PNAS, 103(23), 8577-8582.
- [Palla et al. 2005] G. Palla, I. Derenyi, I. Farkas, and T. Vicsek (2005) Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society. Nature, 435(7043), 814-818
- [Pool et al. 2014] S. Pool, F. Bonchi, M. van Leeuwen (2014) Description-driven Community Detection, Transactions on Intelligent Systems and Technology 5 (2)
- [Psorakis et al. 2011] I. Psorakis, S. Roberts, M. Ebden, and B. Sheldon. Overlapping Community Detection using Bayesian Non-Negative Matrix Factorization. Phys. Rev. E 83, 066114
- [Puppe et al. 2008] F. Puppe, M. Atzmueller, G. Buscher, M. Huettig, H. Lührs, and H.-P. Buscher (2008) Application and Evaluation of a Medical Knowledge-System in Sonography (SonoConsult). In Proc. 18th European Conference on Artificial Intelligence (ECAI 20008), pp. 683-687
- [Ruan et al. 2013] Y. Ruan, D. Fuhry, and S. Parthasarathy (2013). Efficient Community Detection in Large Networks Using Content and Links. Proc. 22nd International Conference on World Wide Web, pp. 1089–1098, ACM.

References (cont.)

- [Tang & Liu 2010] L. Tang and H. Liu (2010) Community Detection and Mining in Social Media. Synthesis Lectures on Data Mining and Knowledge Discovery, 2(1), 1-137. Morgan & Claypool Publishers
- [Steinhaeuser & Chawla 2008] K. Steinhaeuser, N. V. Chawla (2008) Community Detection in a Large Real-World Social Network. Social Computing, Behavioral Modeling, and Prediction, pp. 168–175, Springer
- [Silva et al. 2012] A. Silva, W. Meira Jr., and M. J. Zaki (2010) Structural Correlation Pattern Mining for Large Graphs. Proc. Workshop on Mining and Learning with Graphs. MLG '10, pp. 119–126. New York, NY, USA: ACM.
- [Smith et al. 2014] L. M. Smith, L. Zhu, K. Lerman, and A. G. Percus. Partitioning Networks with Node Attributes by Compressing Information Flow. arXiv:1405.4332
- [Scholz et al. 2013] C. Scholz, M. Atzmueller, A. Barrat, C. Cattuto, and G. Stumme (2013). New Insights and Methods For Predicting Face-To-Face
- Contacts. Proc. 7th Intl. AAAI Conference on Weblogs and Social Media, Palo Alto, CA, USA, AAAI Press.
- [Wassermann & Faust 1994] S. Wasserman, and K. Faust (1994) Social Network Analysis: Methods and Applications. Structural Analysis in the Social Sciences. Cambridge University Press, 1 edition.
- [Wrobel 1997] S. Wrobel (1997) An Algorithm for Multi-Relational Discovery of Subgroups. In Proc. 1st Europ. Symp. Principles of Data Mining and Knowledge Discovery, pages 78–87, Heidelberg, Germany. Springer Verlag.

References (cont.)

- [Xie et al. 2013] J. Xie, S. Kelley, and B. K. Szymanski (2013) Overlapping Community Detection in Networks: The State-of-the-art and Comparative Study. ACM Comput. Surv., 45(4):43:1–43:35.

- [Xu et al. 2012] Z. Xu, Y. Ke, Y. Wang, H. Cheng, and J. Cheng (2012) A Model-based Approach to Attributed Graph Clustering. Proc. ACM International Conference on Management of Data. SIGMOD '12, pp. 505–516, New York, NY, USA. ACM.

- [Yang et al. 2013] J. Yang, J. McAuley, and J. Leskovec (2013) Community Detection in Networks with Node Attributes. Proc. IEEE International Conference on Data Mining (ICDM), pp. 1151–1156. IEEE Press, Washington, DC, USA

- [Zachary, 1977] W. W. Zachary (1977) An Information Flow Model for Conflict and Fission in Small Groups. Journal of Anthropological Research, 452-473.

- [Zhou et al. 2009] Y. Zhou, H. Cheng, and J. X. Yu (2009) Graph Clustering Based on Structural/Attribute Similarities. Proc. VLDB Endow., 2(1), 718–729.