Overview of the Lecture

Part 1: Information retrieval

- Natural language as mechanism to share information

Part 2: Mining unstructured data

- Inferring structure from data aggregated from distributed sources
- Social graph mining (clustering), Document classification, Recommender systems (prediction), Association rule mining

Part 3: Knowledge graphs

- Creating and using shared formal models

Part 2: Data Mining

2. Data Mining - Overview

2.1 Introduction to Data Mining

2.2 Mining Social Graphs
- 2.2.1 Louvain Modularity Algorithm
- 2.2.2 Girvan-Newman Algorithm

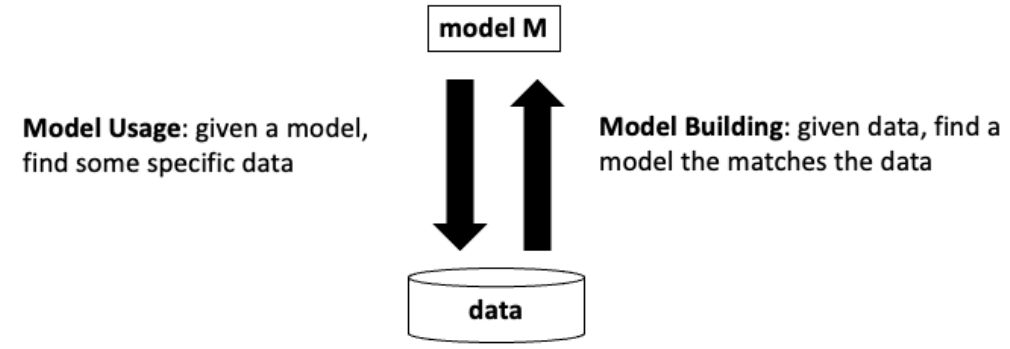
2.3 Document Classification
- 2.3.1 kNN
- 2.3.2 Naïve Bayes Classifier
- 2.3.3 Classification Using Word Embeddings
- 2.3.4 Transformer Models

2.4 Recommender Systems
- 2.4.1 Collaborative Filtering
- 2.4.2 Content-based Recommendation
- 2.4.3 Matrix Factorization
- 2.4.4 SLIM, Sparse Linear Methods
- 2.4.5 Evaluation of Recommender Systems

2.5 Association Rule Mining
- 2.5.1 Association Rules
- 2.5.2 Scoring Function
- 2.5.3 Apriori Algorithm
- 2.5.4 FP Growth

2.1 INTRODUCTION TO DATA MINING

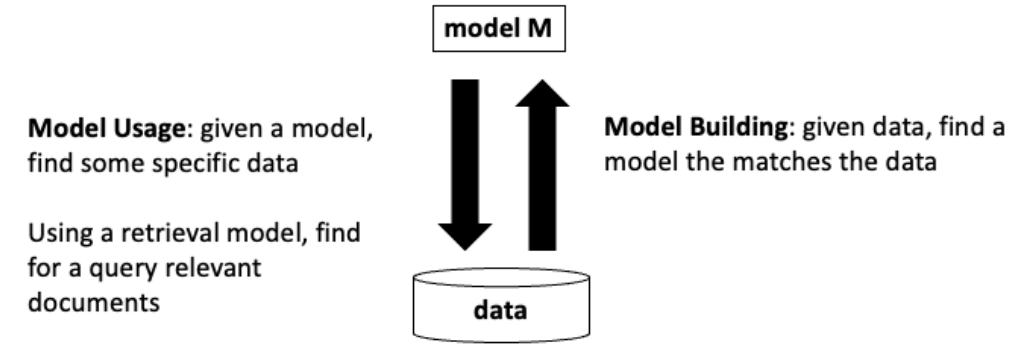
Information Management Tasks



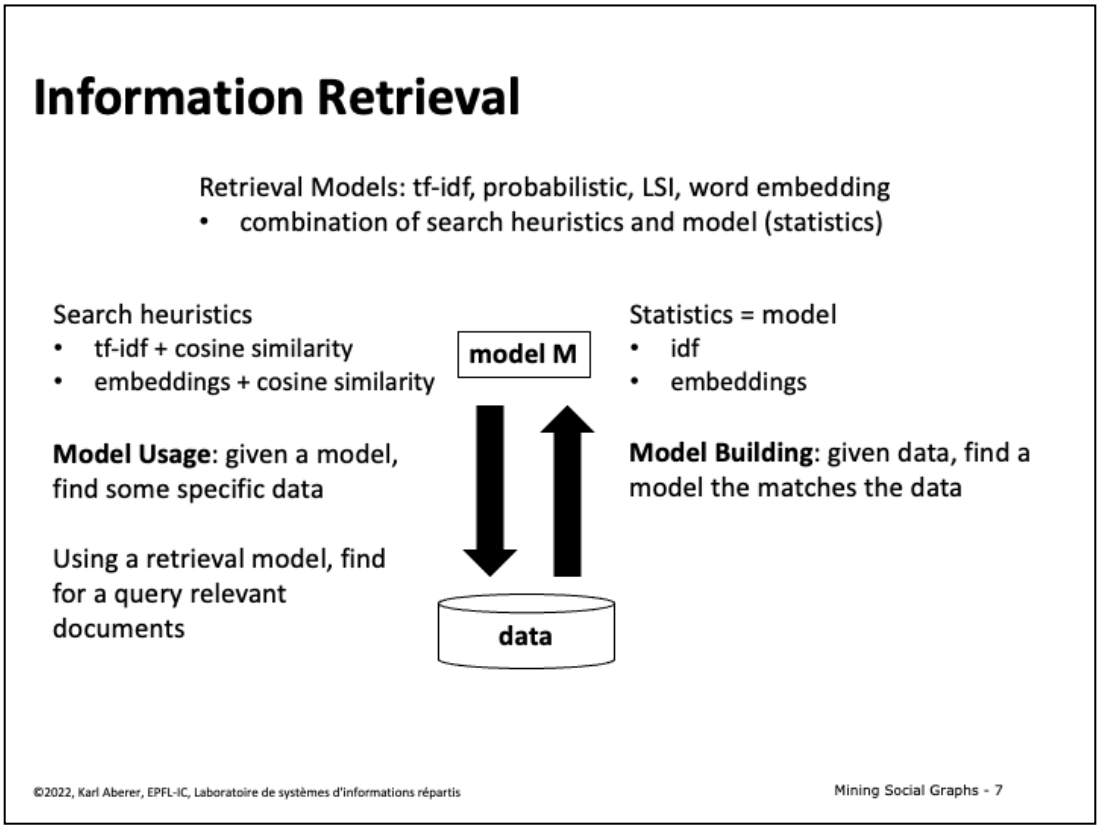
The basic problem of information management is to provide meaningful models for data, and to use them for given tasks.

Information Retrieval

Rertrieval Models: tf-idf, probabilistic, LSI, word embedding



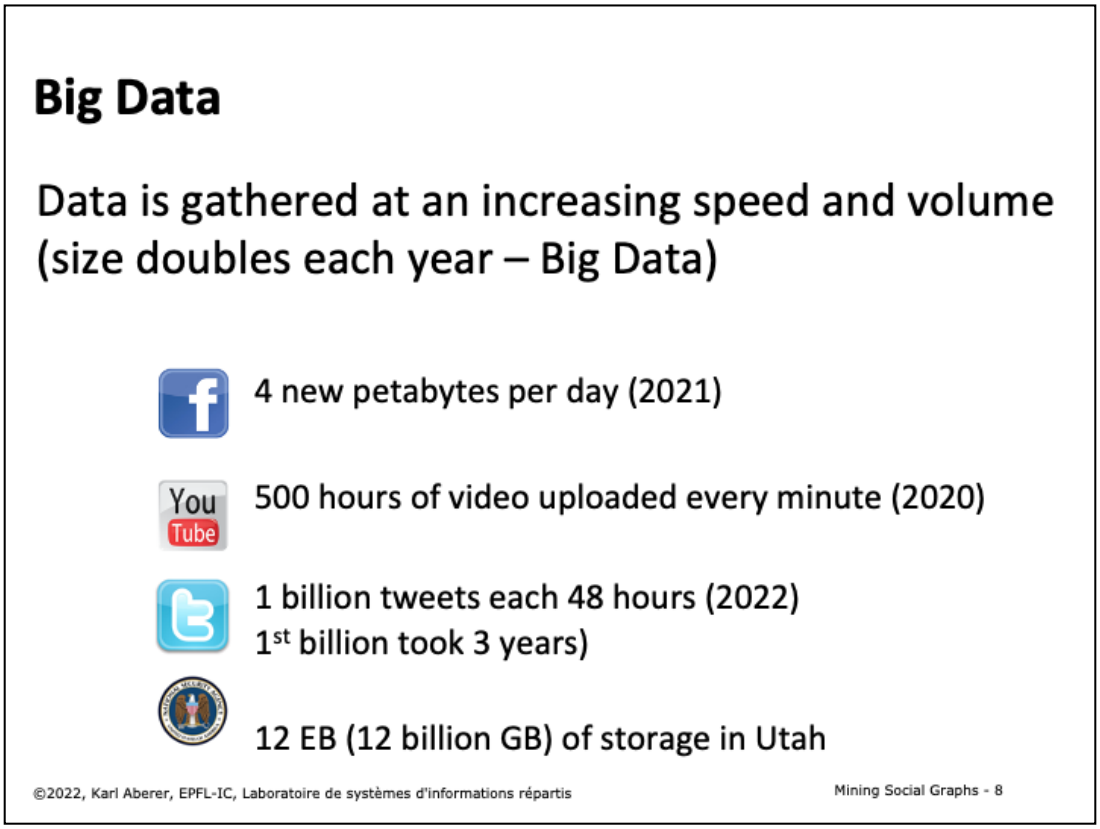
For information retrieval, we have proposed various so-called retrieval models. The purpose of those models was to support a specific task, the retrieval of meaningful documents for a given query or information need of a user.



If we inspect the retrieval models in more detail, we observe that they are a combination of two components:

1. A model derived from the data, which can be some very simple statistics, like idf, or the outcome of a more complex learning algorithm, such as word embeddings.
2. Some heuristics to perform the given task of search. In the context of information retrieval these are generally based on computing from the model and characteristics of the individual documents some similarity metrics, e.g., using cosine similarity and producing then based in it a result for the user.

Therefore we can consider information retrieval as a combination of learning a model from the data, and then using this model to perform a specific task. In the following we will study the problem of learning models in a more general context, used for tasks that are different from information retrieval and applied to data that is different from text.



The growing use of information technology in many domains produces rapidly growing amounts of data in business, science and on the Web. Recently the rate at which digital data is produced started exceeding the amount of storage that can be provided.

These data collections potentially contain a wealth of hidden information, that needs to be discovered. Businesses can learn from their transaction data more about the behavior of their customers and therefore can become more efficient by exploiting this knowledge. Science can obtain from observational data (e.g., satellite data, sensor data) new insights on scientific problems. Web usage data can be analyzed and used to optimize information access, but as well to implement novel business models, such as for advertising on the Web.

The task of extracting useful information from large datasets is called **data mining** and has in the recent years become one of the most dynamically evolving areas in computer science in general.

Data Mining

Most of this data is useless


pictures of drunk people

videos of cats


royal baby, Justin Bieber


SMS to girlfriend

Having increasing amounts of data does not necessarily imply having more useful information.

Data Mining Tasks

But some are interesting


political orientation of people

acts of violence


opinions on products


SMS by a president

However, some of the information could be very interesting.

Data Mining Challenge

Transform masses of data into actionable intelligence

- From transaction data to market insights
- From observational data (satellite, sensors) to new scientific hypothesis
- From web usage data to better user interfaces
- ...

Analysis of large data sets to find relationships and to summarize the data in novel ways that are both understandable and useful

The data mining challenging is to extract from large amounts of data, the information that is relevant for performing a given task. That is called actionable intelligence (but has many other names as well, such as business intelligence, market intelligence, open source intelligence, technology intelligence etc). For information to be useful it has to satisfy the following two key criteria:

- Understandable: a human should be able to interpret the insights to take decisions
- Useful: the insights should help to take decision that have practical impact and utility; this also implies that insights that are not « surprising » but just reproduce existing knowledge are also not very useful.

Tackling the Data Mining Challenge

Practical Questions

- Data Access (ownership!)
- Domain knowledge (expertise!)

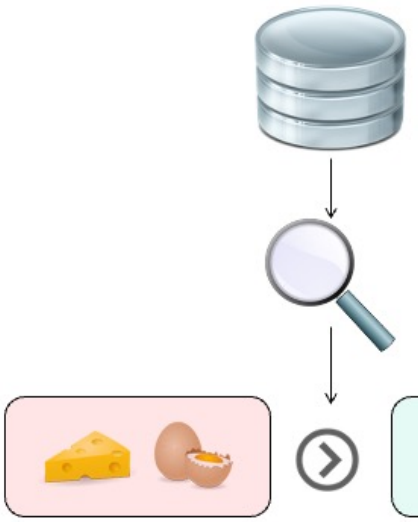
Technical Questions

- Data mining: algorithmic approaches to produce insights
- (Big) Data management (store, index, retrieve)

The challenges to obtain insights with data mining are numerous. The following are some of the key challenges:

1. The data: often massive amounts of data exist, but are often not easily accessible, protected for legal, organizational, economic or political reasons. Also the data can be distributed over many different systems that are not very accessible.
2. The questions: searching for insights into data requires to have at least a general idea of what we are looking for, or what could be an interesting and useful insight. Without understanding the objective, it is hard to find useful answers.
3. The algorithms: Extracting interesting information out of data, requires efficient and smart algorithms. This is what we will study in the following. Once we understand and master such algorithms, the real challenges will be point 1 and 2.
4. Systems for handling Big Data: this is an area that has made huge progress in the recent years, in particular due to the needs of the big Internet companies. Many of the tools are now available as open source and within the cloud.

Example: Shopping Basket Analysis



Purchase transactions

Wal-Mart

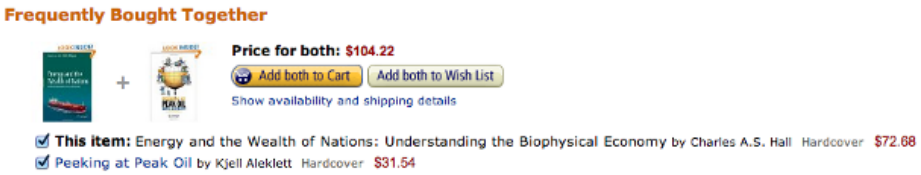
- On Fri afternoon 60% of people who buy diapers also buy beer
- Move beer close to diapers to sell more beers!

Association rule mining

A classical example of a data mining problem is ”shopping basket analysis". Retail stores gather information on what items are purchased by their customers. The expectation is, by finding out what products are frequently purchased together, identifying ways to optimize the sales of the products by better targeting certain groups of customers (e.g., by planning the layout of a store or planning advertisements). A well-known example was the (presumable) discovery that people who buy diapers also frequently buy beers (probably exhausted fathers of small children). Therefore, nowadays one finds frequently beer close to diapers in supermarkets, and of course also chips close to beer. Similarly, amazon exploits this type of associations in order to propose to their customers books that are likely to match their interests. Searching for methods for address this type of problem was resulting one of the best-known data mining techniques: association rule mining.

Other Uses for Association Rules

Amazon



Analysis of Query Logs (Query Expansion)

- Users that search for "Obama President" often also search for "Obama President Elections"

The original term of "shopping basket analysis" does not imply that the resulting methods are limited to analyzing shopping behaviors. The same techniques can and have been applied in many other contexts, including recommender systems, search systems, text analysis, query log analysis etc.

Classes of Data Mining Problems

Local properties

- Patterns that apply to part of the data
  - e.g., buy diapers ➔ buy beers

Global model

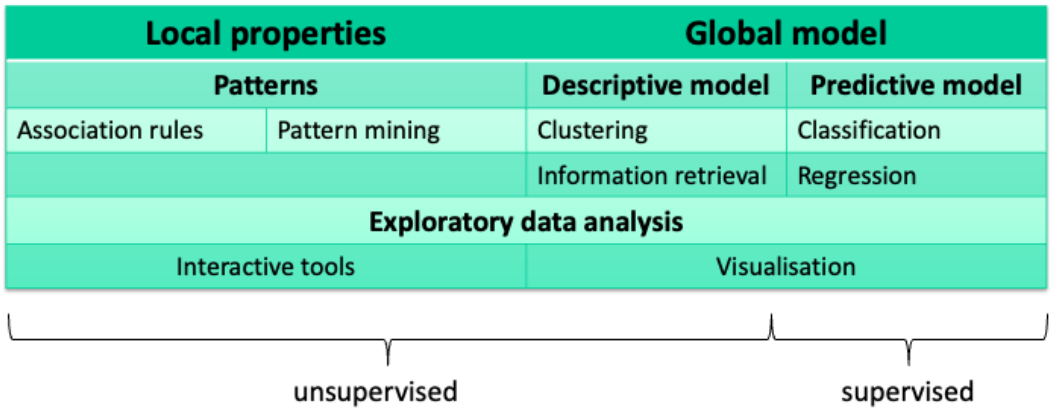
- Descriptive **structure** of the data
  - e.g., 3 types of customer behaviour
- Predictive **function** of the data
  - e.g., dist(beer, diapers)=1 ➔ +10% beer sales

Association rule mining is just one example of a data mining algorithm. Data mining algorithms can in general by classified according the type of models they provide.

A first distinction is made among data mining algorithm that identify global structures of the data, either in the form of summaries of the data or as globally applicable rules, and data mining algorithms that provide models that apply only locally, i.e., to some subset of the data, as patterns or outliers found in the data. The example of association rule mining is a typical case of discovering local patterns.

Data mining algorithms for finding global models of the data can be further distinguished into approaches that are used to characterize the data by identifying some global structure, and into techniques that allow to make predictions on data that has not yet been seen.

Data Mining Overview

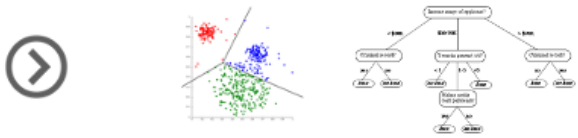


This graphic shows a basic categorization of different data mining methods. In this categorization we can interpret information retrieval as the use of descriptive models. In addition to purely algorithmic methods, data mining comprises also interactive approaches for exploratory data analysis. This is particularly interesting when the user has no good idea of what data to expect.

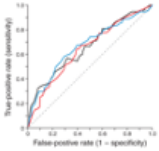
Another important categorization of data mining methods is the distinction of unsupervised and supervised methods, as in general for machine learning. In supervised methods the algorithms are provided with data that provides the intended outputs for prediction tasks.

Components of Data Mining Algorithms

1. Pattern structure/model representation
   - What we look for?



2. Scoring function
   - How well the model fits the data set?



Each data mining algorithm can be characterized by four components:

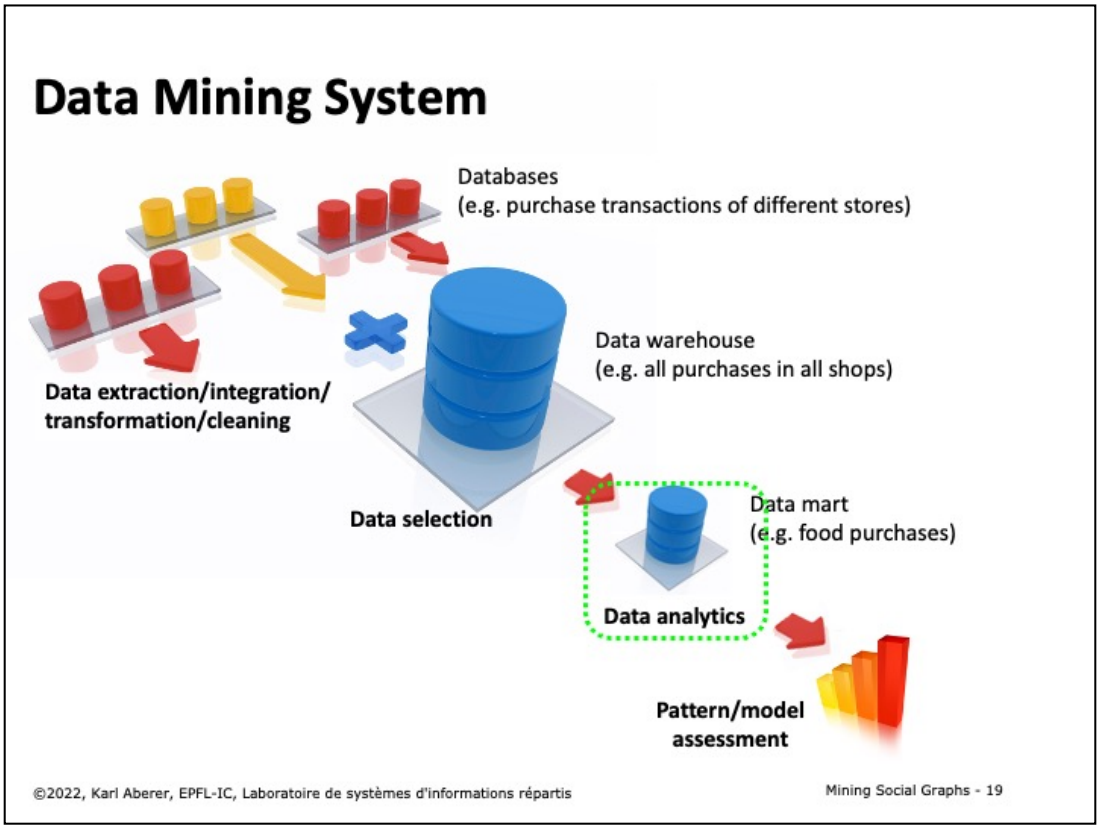
- The models or patterns that are used to describe what is searched for in the data set. Typical examples of models are dependencies, clusters and decision trees.
- The scoring functions that are used to determine how well a given data set fits the model. In information retrieval these were the different measures used to evaluate the quality of a method, like recall or precision.
- The method that is applied in order to identify models from the data set that score well with respect to the scoring function. This requires efficient search algorithms to quickly limit the relevant models based on analyzing the data.
- Finally, the scalable implementation of the methods for large data sets. Here indexing techniques and efficient secondary storage management are typically required. This corresponds to the use of inverted files in information retrieval.

Components of Data Mining Algorithms

3. Optimisation and search
   - How to tune the parameters of the model? [opt]
   - How to find data satisfying a pattern? [search]

4. Data management
   - How to implement the algorithm for very large data sets?

In particular, the aspects of optimization of data search and efficient data management of large dataset delimit data mining from related areas such as statistics and machine learning: data mining algorithms can be understood as statistical or machine learning techniques that scale well to large data sets.



Data mining algorithms are part of data mining system that support the pre-processing of the date and post-processing of the results. A data mining system performs the following typical tasks:

- First, the data needs to be collected from the available data sources. Since these data sources can be distributed and heterogeneous databases, database integration techniques are applied. The integrated data is kept in so-called data warehouses or data lakes, databases replicating and consolidating the data from the various data sources. An important concern when integrating the data sources is data cleaning, i.e., removing inconsistent and faulty data as far as possible. The tasks of data integration and data cleaning are supported by so-called data warehousing systems.
- Once the data is consolidated in a data warehouse, subsets of the data can be selected from the data warehouse for performing specific data mining tasks, i.e., tasks targeting a specific question. This task-specific data collections are sometimes called data-marts. The data-mart is the database to which the specific data mining algorithm is applied.
- The data mining task is the process of detecting interesting patterns in the data. This is what generally is understood as data mining in the narrow sense. We will introduce in the following examples some of the most common techniques that are used to perform this task (e.g., association rule mining).
- Once specific patterns are detected they can be further processed. Further processing may include the evaluation of the "interestingness" of patterns for the specific problem at hand and the implementation of actions to react on the discovery of a pattern.

Each of the steps described can influence the other steps. For example, patterns or outliers detected during data mining may indicate the presence of erroneous data, rather than interesting features in the source databases. This may imply adaptations of the data cleaning process during data integration.

Data Mining ≠ Machine Learning

What is in common

- In DM for data analytics frequently typical ML methods are used, though not always (e.g., visual mining, simple statistics)

What is different

- Data: DM is always applied to large datasets, ML not (e.g., reinforcement learning)
- Scope: DM comprises of the whole process of data integration, cleaning, analysis
- Goal: DM aims at detecting unsuspected patterns, ML may have other goals (e.g., winning a game)

Often data mining and machine learning are used like synonyms. Though both are exploiting the same algorithmic techniques, they can have different scope and goals. In particular, data mining is specifically targeted at analysis of large datasets and concerned with the resulting performance questions.

Detecting unusual behaviours in network logs is typically done by using

A. local rule discovery
B. predictive modelling
C. descriptive modelling
D. exploratory data analysis
E. all of these methods can be applied

2.2 MINING SOCIAL GRAPHS

Mining Graphs

Data Mining can be performed on different types of data

- Structured data: tables, graphs
- Unstructured data: text, images, sensor data

Graph data is increasingly important

- Social networks and Web
- Knowledge Graphs
- Scientific data on networks
- Graph structure can also be inferred from distance measures (e.g., for documents)

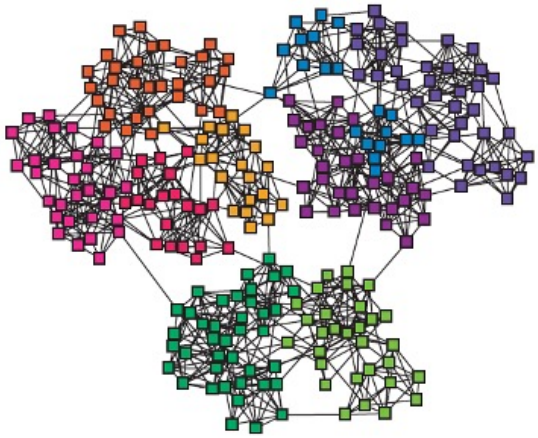
Data mining is applied for a wide range of both structured and unstructured data. It has traditionally evolved from analyzing large transactional databases, typically for business applications,. Structured data graph data is playing an increasingly important role in data mining, due to a growing number of graph data sources. One area where graph data play an obvious role is in social networks, where social interactions and relationships are modelled as graphs. This availability of graph data and the need to understand the structure of large graphs have been driving factors in the development and wide-spread application of graph mining algorithms.

It is also worthwhile to note that graph mining algorithms can be applied to graphs that are generated from other data types. For example, document similarity measures based on text embeddings could be used to generate graph structures for document collections.

Graphs and Clustering

Graphs often contain structure

– Clusters (also called communities, modules)

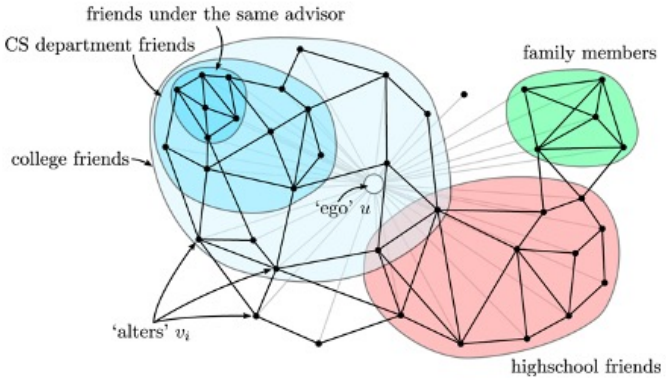


It is widely observed that natural networks contain structure. This is true for social networks (e.g., social media platforms, citation networks), as well as for many natural networks as we find them in biology. Graph-based clustering aims at uncovering such hidden structures.

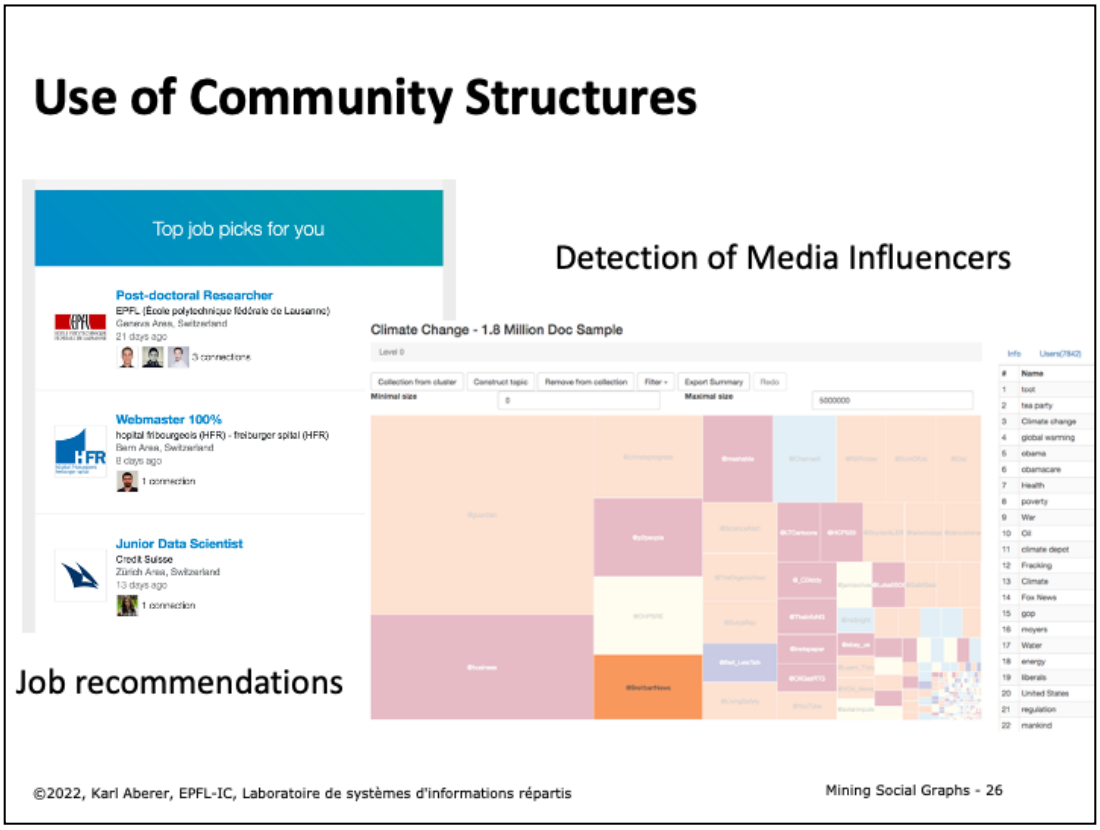
Social Network Analysis

Clusters (communities) in social networks (Twitter, Facebook) related to

- Interests
- Level of trust



In social networks mining community structures is particularly popular. Consider the social network neighborhood of a particular social network user, i.e., all other social network accounts that are connected to the user, either through explicit relationships, such as a follower graph, or through interactions such as likes or retweets. Typically, the social neighborhood would decompose into different groups, depending on interests and social contexts. For example, the friends from high school would have a much higher propensity to connect to each other, than with members of the family of the user. Thus, they are likely to form a cluster. Similarly, other groups with shared interest or high mutual level of trust would form communities. Graph mining is a tool to detect these types of communities.



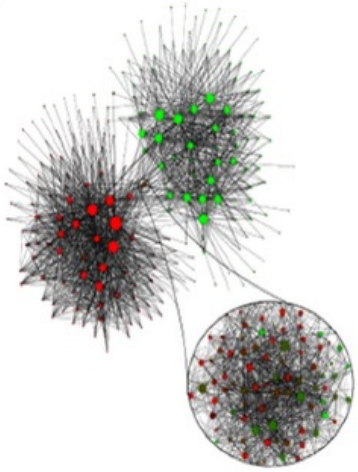
Many tasks can benefit from a reliable community detection algorithm. Online Social Networks rely on their underlying graph to recommend content, for example relevant jobs. Knowing to which community a user belongs can improve dramatically the quality of such recommendations.

Another typical use of community detection is to identify media influencers. Community detection can first be used to detect communities that share in social networks common interests or beliefs (e.g., in the discussion on climate change we might easily distinguish communities that are climate deniers and climate change believers), and then the main influencers of such communities could be identified.

Use of Community Structures: Social Science

Call patterns in Belgium cell phone network

– Two almost separate communities



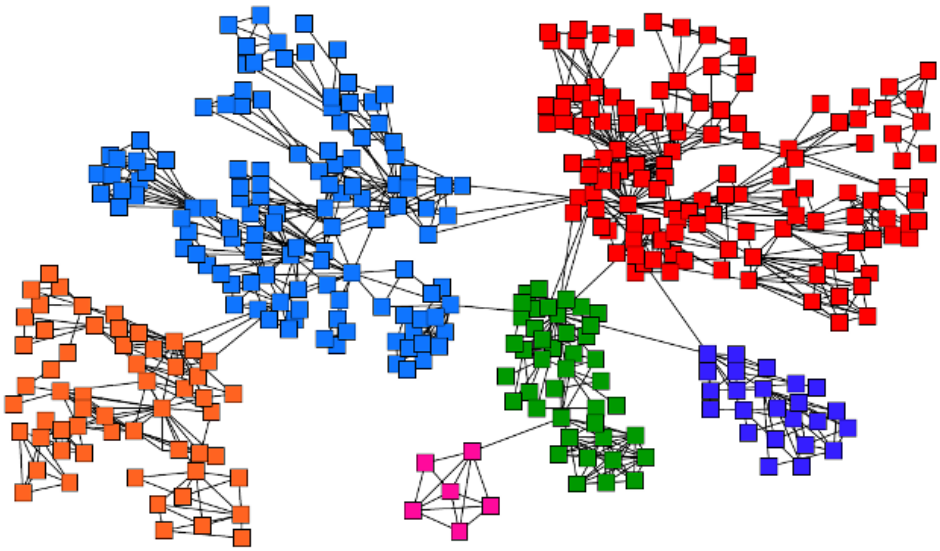
V.D. Blondel et al, *J. Stat. Mech.* P10008 (2008).

In 2008 Vincent Blondel and his students have started applying a new community detection algorithm to the call patterns of one of the largest cell phone operators in Belgium. It was designed to identify groups of individuals who regularly talk with *each other* on the phone, breaking the whole country into numerous small and not so small communities by placing individuals next to their friends, family members, colleagues, neighbors, whom they regularly called on their cell phone. The result was somewhat unexpected: it indicated that Belgium is broken into two huge communities, each consisting of many smaller circles of friends. Within each of these two groups the communities had multiple links to each other. Yet, these communities never talked with the communities in the other group (guess why?). Between these two large groups we find a third, much smaller group of individuals, apparently mediating between the two parts of Belgium.

Which of the following graph analysis techniques do you believe would be most appropriate to identify communities on a social graph?

A. Cliques
B. Random Walks
C. Shortest Paths
D. Association rules

Task: Find Densely Linked Clusters



The intuition behind community detection is that the heavily linked components of the graph belong to the same community. Thus, a community is a set of nodes that has many connections among themselves, but few to other parts of the network. These communities form clusters in the network.

Types of Community Detection Algorithms

Hierarchical clustering

- iteratively identifies groups of nodes with high similarity

Two strategies

- **Agglomerative algorithms** merge nodes and communities with high similarity
- **Divisive algorithms** split communities by removing links that connect nodes with low similarity

In general, community detection algorithms are taking a hierarchical approach. The idea is that in a community smaller sub-communities can be identified, till the network decomposes into individual nodes. In order to produce such a hierarchical clustering, two approaches are possible: either by starting from individual nodes, by merging them into communities, and recursively merge communities into larger communities till no new communities can be formed (agglomerative algorithms), or by decomposing the network into communities, and recursively decompose communities till only individual nodes are left (divisive algorithms).

In the following we will present one representative of each of the two categories of algorithms:

1. The Louvain Algorithm, an agglomerative algorithm
2. The Girvan-Newman algorithm, a divisive algorithm

2.2.1 Louvain Modularity Algorithm

Agglomorative Community Detection

- Based on a measure for community quality (**Modularity**)
- greedy optimization of modularity

Overall algorithm

- **first** small communities are found by optimizing modularity **locally** on all nodes
- then each small community is **grouped** into one new community node
- **Repeat** till no more new communities are formed

The Louvain algorithm is essentially based on the use of a measure, modularity, that allows to assess the quality of a community clustering. The algorithm performs greedy optimization of this measure. The basic idea is straightforward: initially every node is considered as a community. The communities are traversed, and for each community it is tested whether by joining it to a neighboring community, the modularity of the clustering can be improved. This process is repeated till no new communities form anymore.

A short remark on terminology: in mathematics the usual terminology for graph nodes is “vertex”. We will in the following continue to use the term “node” for “vertex”, as it is custom in computer science.

Measuring Community Quality

Communities are sets of nodes with many mutual connections, and much fewer connections to the outside

Modularity measures this quality: the higher the better

$$\sum_{C \in \text{Communities}} (\#\text{edges within } C - \text{expected } \#\text{edges within } C)$$

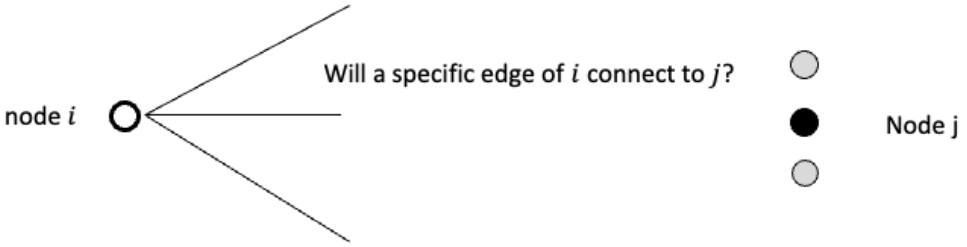
In order to measure the quality of a community clustering, modularity compares the difference between the number of edges that occur within a community with the number of edges that would be expected in a random subset of nodes of the same size and randomly distributed edges.

Expected Number of Edges

Graph with unweighted edges

- m = total number of edges
- k_i = number of outgoing edges of node i (degree)

Observation: there exist $2m$ "edge ends"



distributes k_i edge ends uniformly to other nodes

$\frac{k_j}{2m-1}$ edges end up in node j

Probability of an edge of k_i ending up in node j: $k_i \frac{k_j}{2m-1} \approx \frac{k_i k_j}{2m}$

We now determine the number of edges we expect between nodes with given degrees ki, if edges were purely randomly allocated.

How many edges would we observe if the connections on the graph were generated randomly? To answer this question, we reason as follows: select one of the nodes i with degree ki. What is the probability that this edge connects to another node j with weight kj? If there are a total of m edges in the network, there are 2m edge ends (since each edge ends in two nodes). If the edge ends are uniformly distributed and there are 2m-1 remaining edge ends in the graph, the probability to connect exactly to node j would be kj/2m-1. Applying the same argument to all outgoing edges of node i, node j will connect to node i with probability ki*(kj/2m-1). For large m we can replace 2m-1 by 2m.

Modularity

Modularity measure Q

- A_{ij} = effective number of edges between nodes i and j
- C_i, C_j = clusters of nodes i and j

$$Q = \frac{1}{2m}\sum_{i,j}\left(A_{ij} - \frac{k_i k_j}{2m}\right)\delta(C_i, C_j)$$

Expected number of edges

Effective number of edges

Properties

- Q in [-1,1]
- 0.3-0.7 < Q means significant community structure

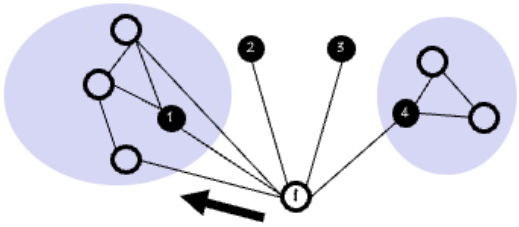
Given the expected number of edges we can now formulate the modularity measure as the total difference of number of expected and effective edges for all pairs of nodes from the same cluster. The delta function assures that only nodes belonging to the same cluster are considered, since it returns 1 when ci = cj.

Due to normalization the measure returns values between -1 and 1. In general, if modularity exceeds a certain threshold (0.3 to 0.7) the clustering of the network is considered to exhibit a good community structure.

Locally Optimizing Communities

What is the modularity gain by moving node i to the communities of any of its neighbors?

– Test all possibilities and choose the one that increases modularity the most, if it exists

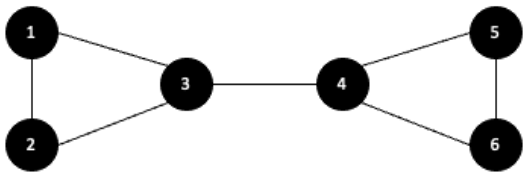


Given the modularity measure, the next question is now how to use it to infer the communities. This is performed through local optimization. The algorithm sequentially traverses all nodes of the network, and for each node checks how the modularity can be increased maximally by having the node joining the node of a neighboring community. If such a neighboring node exists, the node will join the community.

Example

Initial modularity Q = 0

Start processing nodes in order



We illustrate the algorithm for a simple example. Initially, the modularity is zero since all nodes belong to different communities. We start now to process the nodes in some given order, e.g., by increasing identifiers.

Example: Processing Node 1

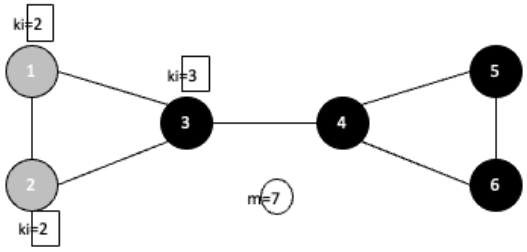
Joining node 1 to node 2

- Q = 1/2*7 (1 – 2*2/2*7) = 1/14 * 10 /14 > 0

Joining node 1 to node 3

- Q = 1/2*7 (1 – 2*3/2*7) = 1/14 * 8/14 > 0

New modularity: 1/14 * 10 /14



Node 1 can either join node 2 or 3, it's two neighbors. To decide which is better, we compute modularity after the join. We see that joining node 2 results in a higher modularity. We can interpret this as follows: since node 3 has more connections, it is more likely to be randomly connected to node 1, thus connecting to node 2 is less likely to be the result of a random connection.

Example: Processing Nodes 2 and 3

Joining node 2 to node 3 (leaving community of node 1)

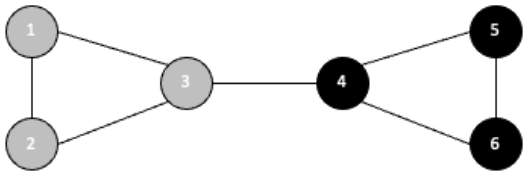
- no improvement

Joining node 3 to community {1,2} (via node 1 or 2)

- $Q = 1/14\ (3 - 4/14 - 6/14 - 6/14) = 1/14 * 26 / 14$

Joining node 3 to node 4

- $Q = 1/14\ 10/14 + 1/14\ (1 - 9/14) = 1/14 * 15/14$



Node 2 will not change the community, as the only alternative would be node 3. But for the same reasons node 1 did not join node 3, also node 2 does not. The next node is node 3: this node has two choices, either to join community {1,2}, or to join node 4. In the first case we obtain one larger community, and in the second case two smaller communities. Computation of modularity reveals that joining join community {1,2}, gives a higher modularity.

Example: Processing Nodes 4, 5 and 6

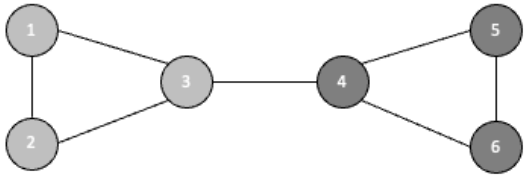
Joining node 4 to community {1,2,3} via 3

- Q = 1/14 (4 - 4/14 - 6/14 - 6/14 – 6/14 – 6/14 - 9/14) = 1/14 * 19/14

Joining node 4 to node 5

- Q = 1/14 * 26/14 + 1/14 * (1 – 6/14) = 1/14 * 34/14

Finally, also node 6 will join the second community

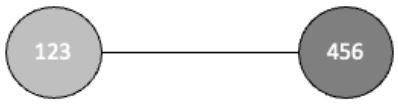


For similar reasons as before, node 4 will join node 5 and finally node 6 will join nodes {4,5}. This completes the first iteration.

Example: Merging Nodes

Now that all nodes have been processed, we merge nodes of the same community in a single new node and restart processing

Will the two remaining nodes merge? Answer: yes



In the next iteration, the current community nodes are collapsed into new nodes representing the communities, and the algorithm is re-run with the new resulting graph. Obviously, now the two remaining nodes will merge into a single community, as the modularity will change from zero to a positive value. Then the algorithm terminates.

Modularity clustering will end up always with a single community at the top level?

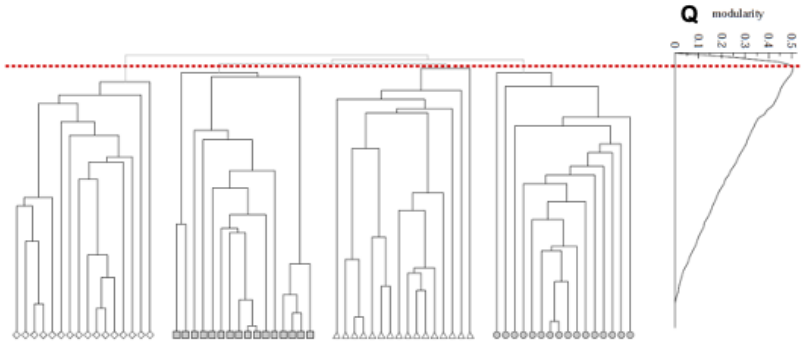
A. true
B. Only for dense graphs
C. Only for connected graphs
D. never

Modularity clustering will end up always with the same community structure?

A. true
B. Only for connected graphs
C. Only for cliques
D. false

Modularity to Evaluate Community Quality

Modularity can also be used to evaluate the best level to cutoff of a hierarchical clustering



Apart from detecting communities, the modularity measure can also be used to evaluate the quality of communities in a hierarchical clustering. This can be done independently of how the clustering has been constructed.

By evaluating modularity at each level of the hierarchy, an optimal modularity value can be determined. This optimal value indicates which are the highest quality communities, and lower levels in the tree hierarchy can be pruned.

Louvain Modularity - Discussion

Widely used in social network analysis and beyond

- Method to extract communities from very large networks very fast

Complexity: O(n log n)

Louvain modularity clustering is widely used method for social network clustering, mainly because of its good computational efficiency. It runs in O(n log n), which makes it applicable for very large networks, for example, for social networks resulting from large Internet platforms, such as social networking sites or messaging services.

2.2.2 Girvan-Newman Algorithm

Divisive Community Detection

- Based on a **betweenness measure** for edges, measuring how well they separate communities
- Decomposition of network by splitting along edges with highest separation capacity

Overall algorithm

- Repeat until no edges are left
  - Calculate betweenness of edges
  - Remove edges with highest betweenness
  - Resulting connected components are communities
- Results in hierarchical decomposition of network

We now introduce a second algorithm for community detection, that belongs to the class of divisive algorithms. This algorithm uses he betweenness measures that quantifies the capacity of an edge to separate the network into distinct subnetworks. Note that while for an agglomerative algorithm we used a measure that quantifies how well communities are connected, for a divisive algorithm we use a measure that quantifies how well communities can be separated.

Using the betweenness measure, the algorithm will recursively split the network into smaller networks, till arriving at individual nodes. This will result in a hierarchical clustering.

Edge Betweenness Centrality

Edge betweenness centrality: fraction of number of shortest paths passing over the edge

$$betweenness(v) = \sum_{x,y} \frac{\sigma_{xy}(v)}{\sigma_{xy}}$$

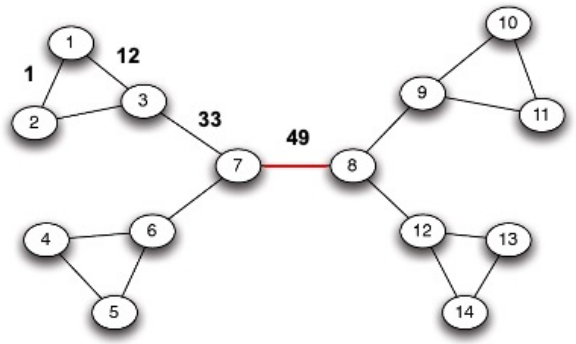
where

σ_{xy}: number of shortest paths from x to y
$\sigma_{xy}(v)$: number of shortest paths from x to y passing through v

Betweenness centrality is an indicator of a node's centrality in a network. It is equal to the number of shortest paths from all nodes to all other nodes that pass through that node. A node with high betweenness centrality has a large influence on the transfer of items through the network, under the assumption that item transfer follows the shortest paths. The concept finds wide application, including computer and social networks, biology, transport and scientific cooperation.

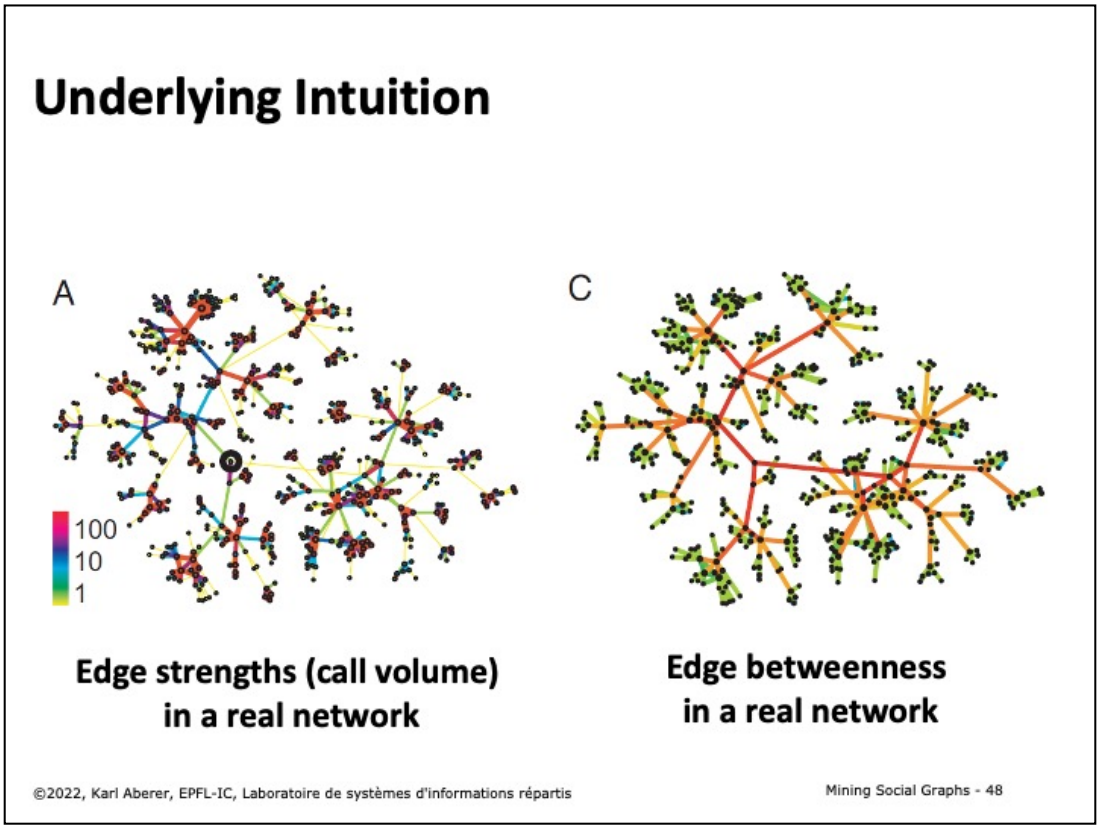
As an alternative measure, one could also consider *random-walk betweenness*. A pair of nodes x and y are chosen at random. A walker starts at x, following each adjacent link with equal probability until it reaches y. Random walk betweenness r_{xy} is the probability that the link $x \rightarrow y$ was crossed by the walker after averaging over all possible choices for the starting nodes x and y. Different to betweenness this measure does consider all paths between the two nodes, not only the shortest paths. However, this measure is expensive to compute.

Example

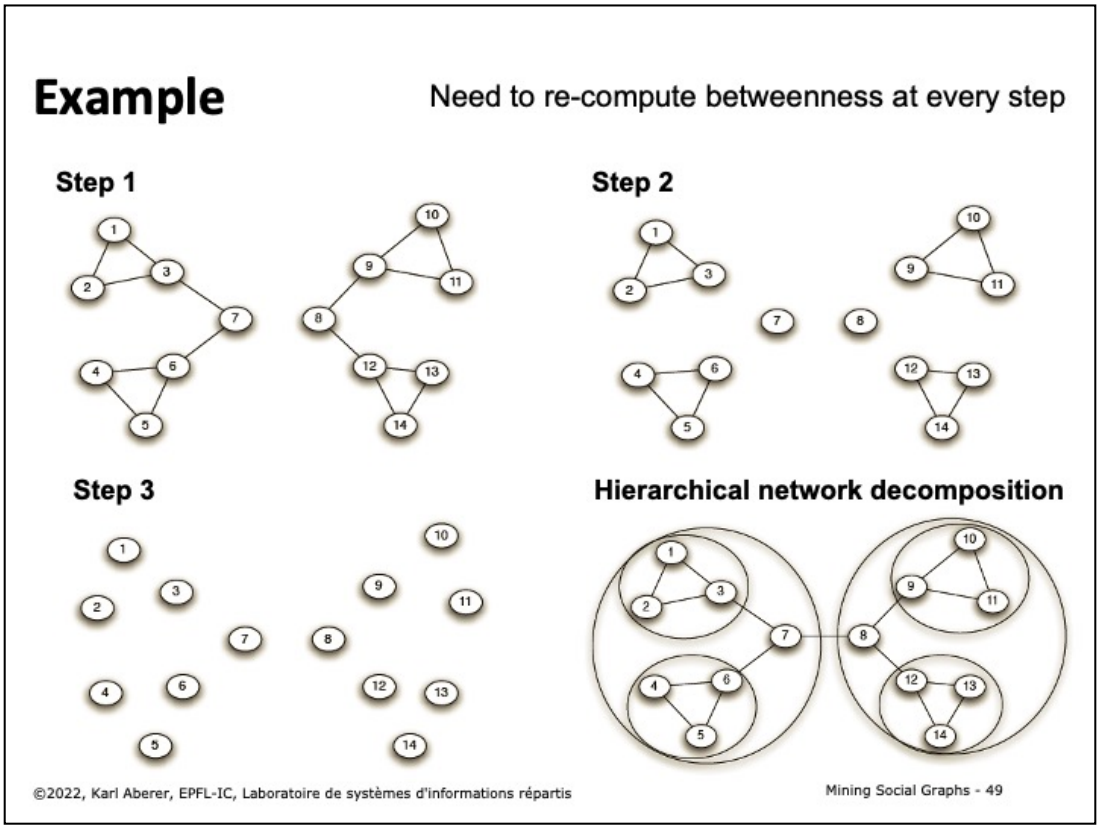


$\sigma_{xy}(v)$

In this example network we illustrate the value of $\sigma_{xy}(v)$ for some selected edges. For example, for the edge 1-2 there is one shortest path between 1 and 2 that traverses the edge 1-2, thus the value is 1. For 1-3 there are shortest paths from the remaining 12 nodes in the network (except node 2) to node 1 that must pass through this edge, thus the betweenness value is 12. For edge 3-7 we have shortest paths reaching both nodes 1 and 2, thus the betweenness value of that edge is significantly higher than of 1-3. For edge 7-8 we have 7 nodes in each of the two subnetworks on the left and on the right (including the nodes 7 and 8). Among each pair of nodes there exists exactly one shortest path. Therefore, the value of $\sigma_{xy}(v)$ is 49.



Betweenness can be considered as a concept that is dual to connectivity. This is illustrated by the two graphs that result from real phone call networks. On the left-hand side, we see the indication of the strengths of connections among the nodes. Communities are tightly connected by such links. On the right-hand side, we see the betweenness measure. Now the links that are connecting different communities have a high strength. This can be interpreted in different ways: on the one hand traffic from one community to another has to traverse these links, on the other hand if such links are cut, communities are separated.

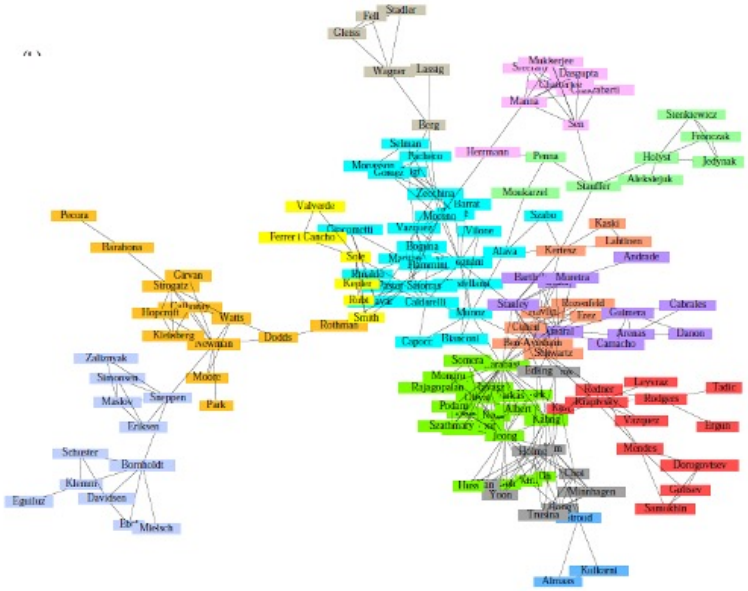


Next, we illustrate the principle of the Girvan-Newman algorithm. It proceeds by recursively removing edges with highest betweenness from the network.

In Step 1 it removes one edge (in the middle) that had the highest betweenness value, resulting in two communities. Next the edges connected to nodes 7 and 8 are removed, and in the third and fourth step the network decomposes completely. By overlaying the communities that have resulted from each step we obtain the final hierarchical clustering.

As the graph structure changes in every step, the betweenness values need to be recomputed in every step. This results in the main cost of the algorithm. We will next explore the question of efficient computation of betweenness.

Girvan-Newman: Sample Results

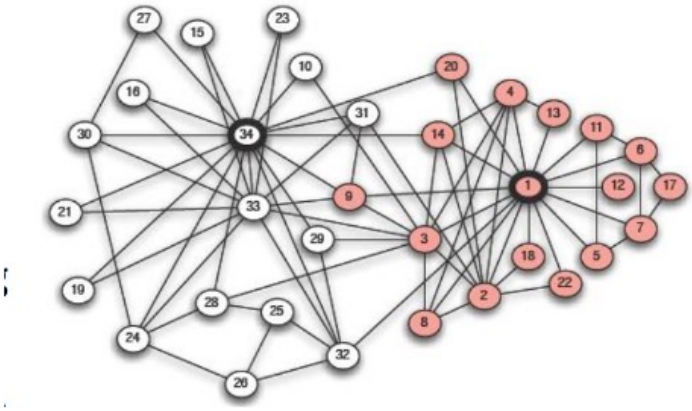


Communities in physics collaborations

The algorithm has been applied in many contexts, due to the computational cost mostly on smaller graphs resulting from social science studies.

Girvan-Newman: Sample Results

Zachary's Karate club: Hierarchical decomposition

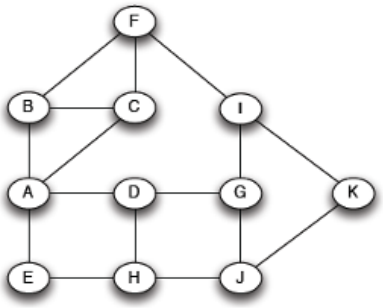


In fact, the origin of the Girvan-Newman algorithm can be traced back to a specific social science study. In this study a network of 34 karate club members was studied by the sociologist Wayne Zachary Links capture interactions between the club members outside the club. The white and gray nodes denote the two communities that resulted after the club decided to split following a feud between the group's president and the coach. The split between the members closely follows the boundaries of the two communities. In the community graph we can identify one outlier, node number 9, where the algorithm wrongly assigns the member at the time of conflict. Node 9 was completing a four-year quest to obtain a black belt, which he could only do with the instructor, node 34.

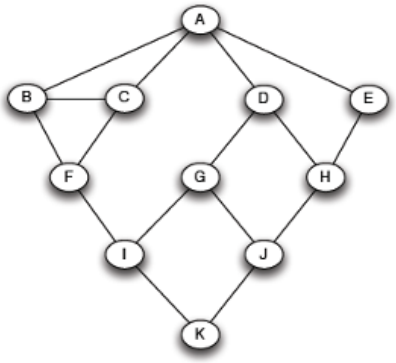
This example has been widely used as a benchmark to test community detection algorithms.

Computing Betweenness - BFS

Computing betweenness of paths starting at node A



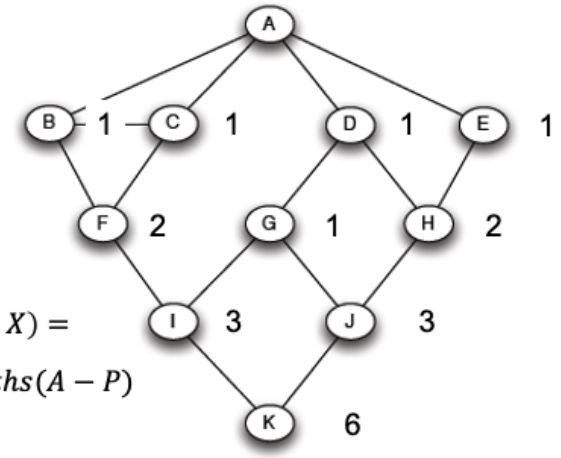
Perform BFS starting from A



We describe now the approach for computing the betweenness values. It is based on a breadth-first search (BFS) starting at every node in the graph. For a given node, e.g., node A, the other nodes are arranged in levels according to increasing distance from the node.

Computing Betweenness – Path Counting

Count the number of shortest paths from A to all other nodes of the network



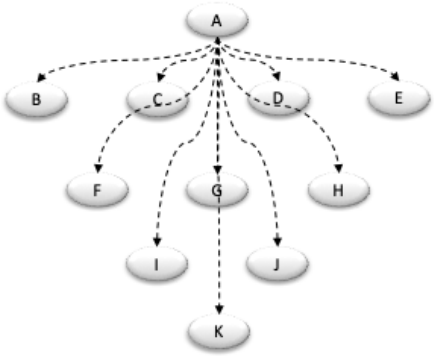
$$\#shortestpaths(A-X) = \sum_{P\ parent\ of\ X} \#shortestpaths(A-P)$$

In a first phase, we count the number of shortest paths that are leading to each node, starting from node A. To do so, the algorithm can at each level reuse the data that has been computed at the previous level. The number of shortest paths leading to a given node corresponds to the sum of number of shortest paths leading to each of its parent needs. For example, the shortest paths leading to node F, are all shortest paths leading to nodes B and C. Note that in this way paths that are not shortest paths are ignored, like the path A-B-C-F.

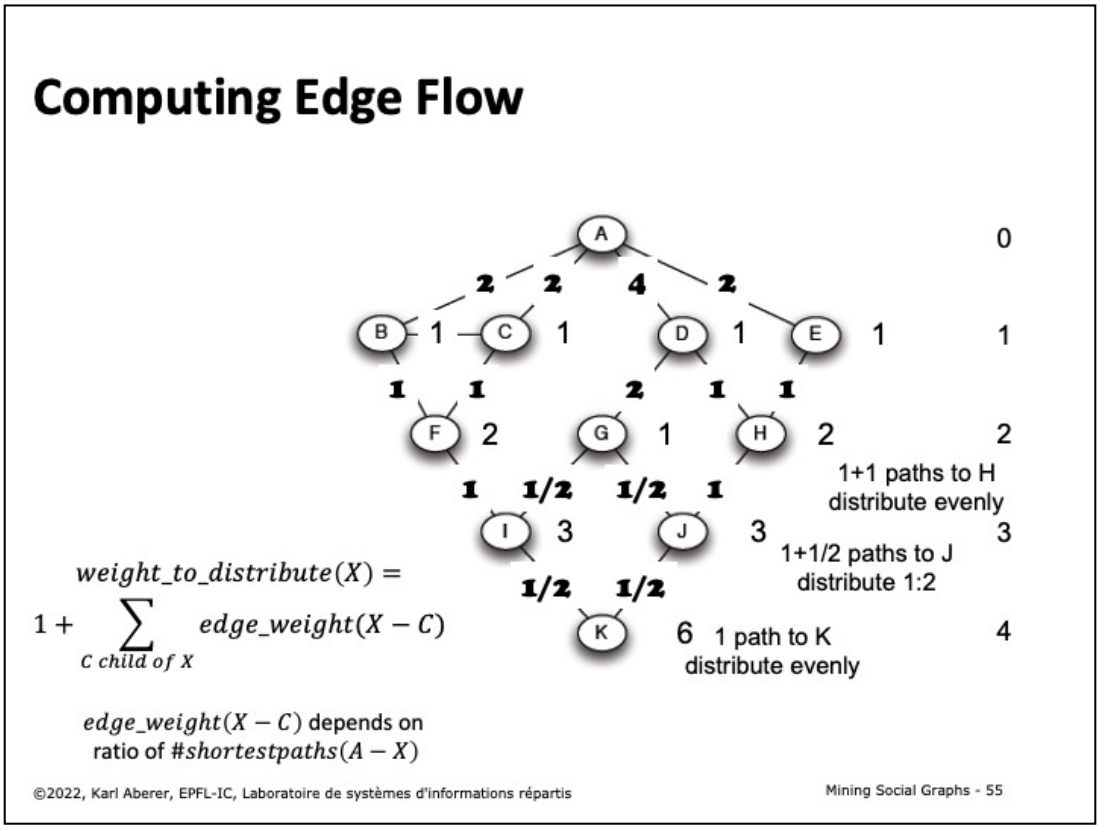
Computing Betweenness – Edge Flow

Edge Flow

- 1 unit of flow from A to each node
- Flow to be distributed evenly over all paths
- Sum of the flows from all nodes though an edge equals the betweenness value



In order to compute the betweenness values for edges, we have to aggregate the contributions of all paths between arbitrary pairs of nodes. To do this the following model is adopted. From each node a flow is emanating to each other node of the graph, which has a total volume of 1. This flow is evenly distributed among all shortest paths, that lead from the source node to the target node. Note that these different shortest paths may be overlapping, i.e., have common edges. Therefore edges that are part of different shortest paths will receive contributions from all the shortest paths passing through them. Finally, the flows will be aggregated for the flows emanating from all nodes, ot obtin the betweenness value of the edge.



Here we illustrate the second phase of the betweenness computation, computing the flow values related to one node, in the example node A. For each node one must consider the flow that is arriving at the node, plus the flows that are passing through the nodes to another target node. For computing the aggregate edge flows, the algorithm will start at the farthest node, in the example node K, which has no shortest paths emanating. Therefore, the only flow it receives is the one assigned to itself, i.e., a flow of 1. Since the node can be reached through different shortest paths, which are represented by the incoming edges from the nodes of the next higher level, this flow is distributed proportionally to the number of shortest paths leading to the parent nodes. In the case of node K, there exist 3 shortest paths leading to its two parent nodes. Therefore, the flows are evenly distributed. This results in flows of ½ for the incoming edges of node K.

For the nodes at the higher levels, the flows destined for the node are summed up with the flows destined for its descendant nodes that have been computed before. For example, node I has an outgoing flow of ½ plus a flow of 1 destined for itself. Therefore, it has to distribute a total flow of 3/2 over its incoming edges. This distribution must be done proportionally to the number of shortest paths arriving at its parent nodes. Since 2 shortest paths arrive at node F and 1 shortest path arrives at node G, the flow of 3/2 is split at a ratio of 2:1 among the two incoming edges of node I. The analogous reasoning is applied to the other nodes.

In general, each node distributes tribute a weight of $1 + \sum_{C\ child\ of\ X} edge\_weight(X - C)$ over the incoming edges, i.e., the flow of 1 arriving at the node plus the sum of the flows on the outgoing edges. The distribution is done according to the ratios of the shortest paths arriving at the parent nodes.

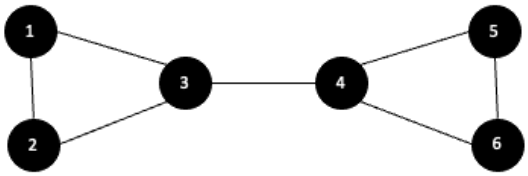
Algorithm for Computing Betweenness

1. Build one BFS structure for each node
2. Determine edge flow values for each edge using the previous procedure
3. Sum up the flow values of each edge in all BFS structures to obtain betweenness value
   - Flows are computed between each pairs of nodes → final values divided by 2

Once the edge flows have been computed for all nodes, the resulting values are summed up. Since for each shortest path two flows have been generated, corresponding to the traversal of the path in the two directions, the final aggregate flow is divided by 2.

$\sigma_{xy}(v)$ of edge 3-4 is …

A. 16
B. 12
C. 9
D. 4



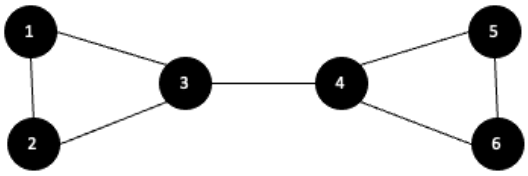
When computing path counts for node 1 with BFS, the count at 6 is …

A. 1

B. 2

C. 3

D. 4



Girvan-Newman Discussion

Classical method

- Works for smaller networks

Complexity

- Computation of betweenness for one link: $O(n^2)$
- Computation of betweenness for all links: $O(L\ n^2)$
- Sparse matrix: $O(n^3)$

The Girvan-Newman algorithm is the classical algorithm for community detection. Its major drawback is its scalability. The flow computation for one link has quadratic cost in the number of nodes, since it is computed for each pair of nodes. If we assume sparse networks, where the number of links is of the same order as the number of nodes, the total cost is cubic. This was also one of the motivations the inspired the development of the modularity-based community detection algorithm.

References

The slides are loosely based also on:

- http://barabasilab.neu.edu/courses/phys5116/

Papers

- Blondel, Vincent D., et al. "Fast unfolding of communities in large networks." *Journal of statistical mechanics: theory and experiment* 2008.10 (2008): P10008
- Girvan, Michelle, and Mark EJ Newman. "Community structure in social and biological networks." *Proceedings of the national academy of sciences* 99.12 (2002): 7821-7826.