Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

RuSSIR
Russian Summer School
in Information Retrieval

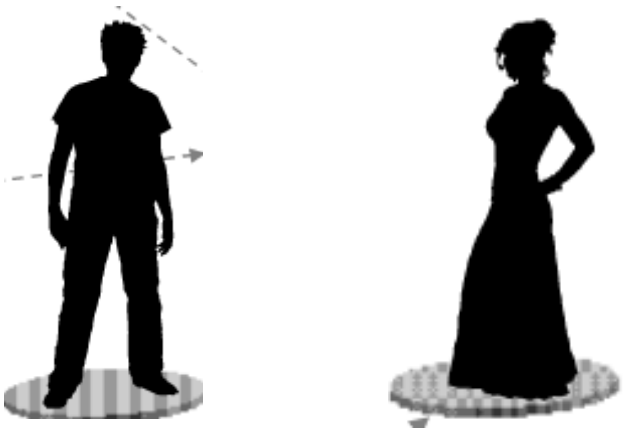
Social Networks, Social Network Analysis and Social Software

Social Networks and Social Network Analysis

- Many social sciences focus on people, organisations, ….
- SNA is the study of “social” relations among people, organisations etc (actors)

Key difference between SNA and other approaches to social science

- **Social sciences usually have focus on attributes of individual actors**

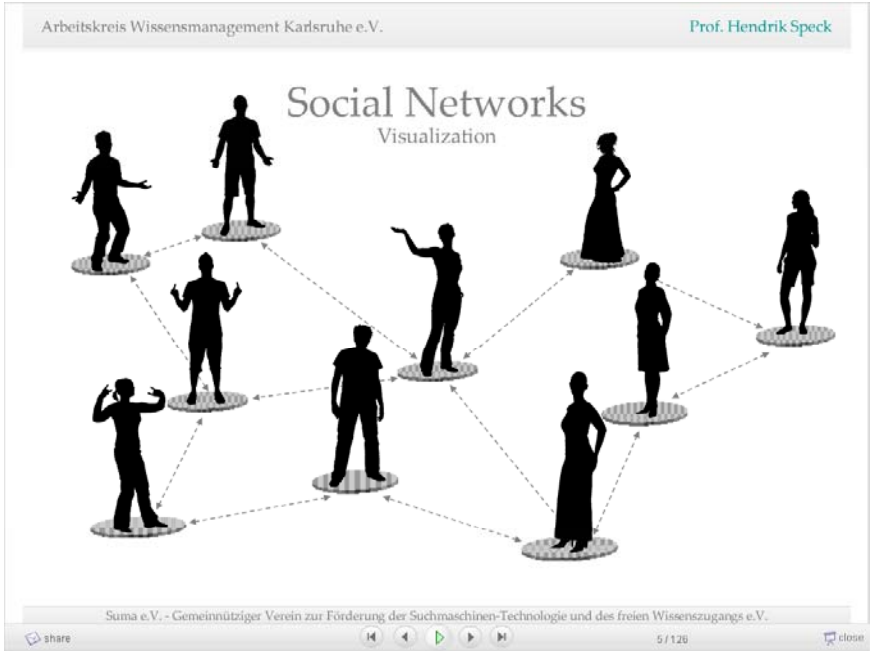


Key difference between SNA and other approaches to social science

- **SNA focus on relationships between actors**

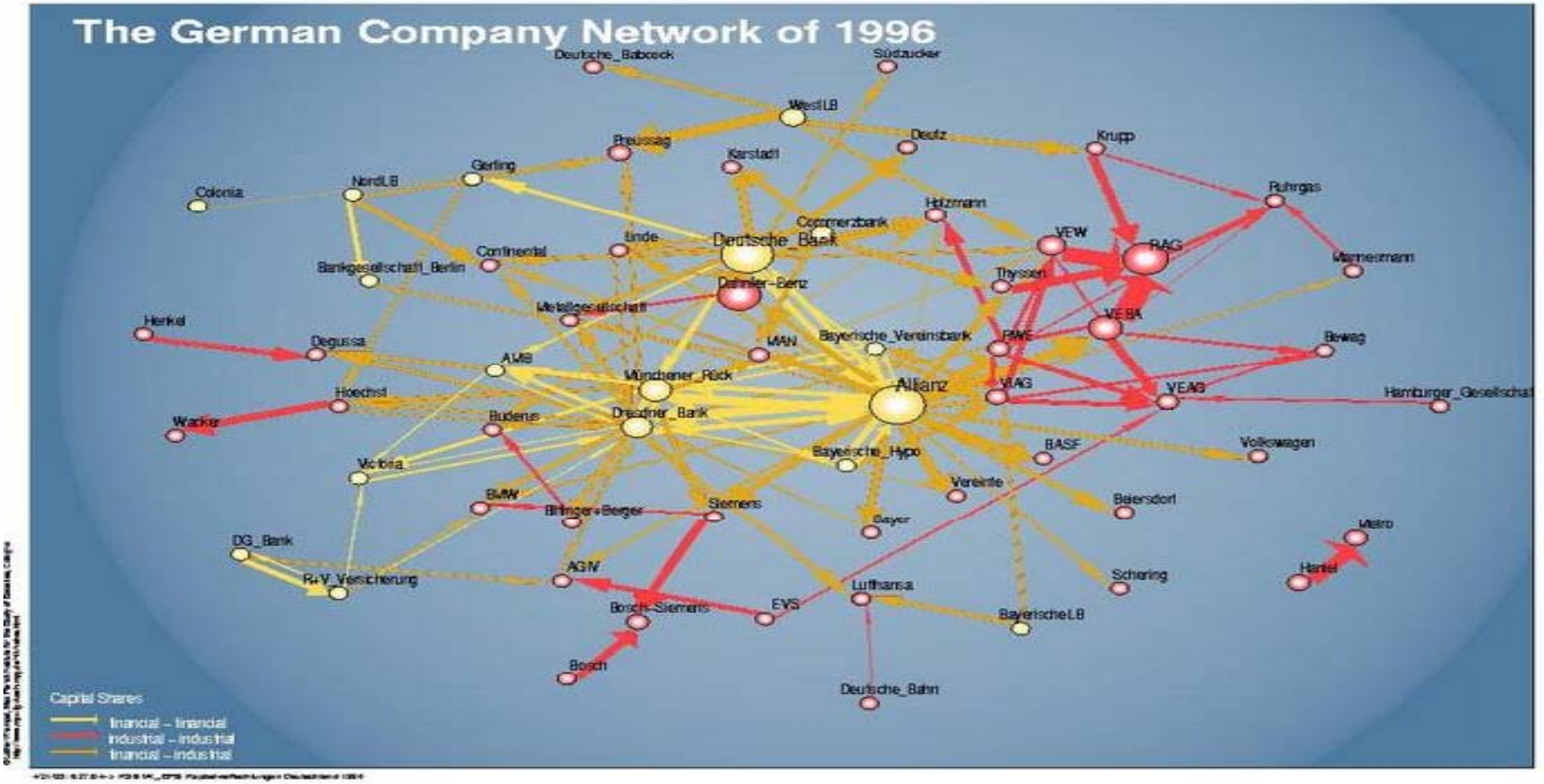
- "Social network analysis reflects a shift from the individualism common in the social sciences towards a structural analysis".
  Garton et al. Studying Online Social Networks

- Sociogram:
  - Jacob Levy Moreno (1889-1974) was a Austrian-American leading psychiatrist and psychosociologist, thinker and educator, the founder of psychodrama, and the foremost pioneer of group psychotherapy.
    Among Moreno's primary contributions to sociometrics was the sociogram. The sociogram is a method of representing individuals as points on graphs and using lines and arcs to represent the relationships between the individuals.



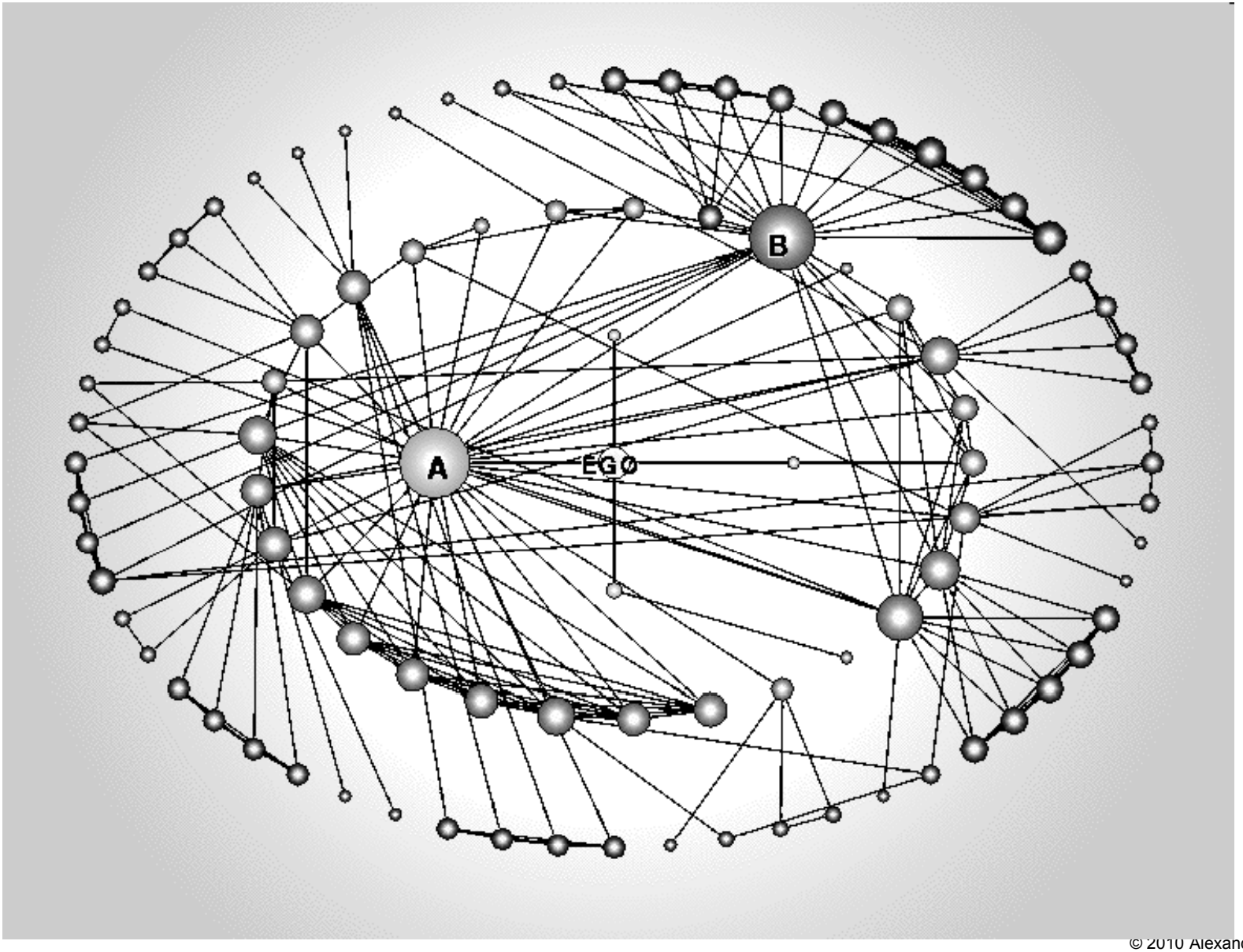
- **Graphics from Prof. Hendrik Speck's tutorial at 5th Karlsruhe Symposium for Knowledge Management in Theory and Praxis, 2007**

Examples of Social Networks – German Companies

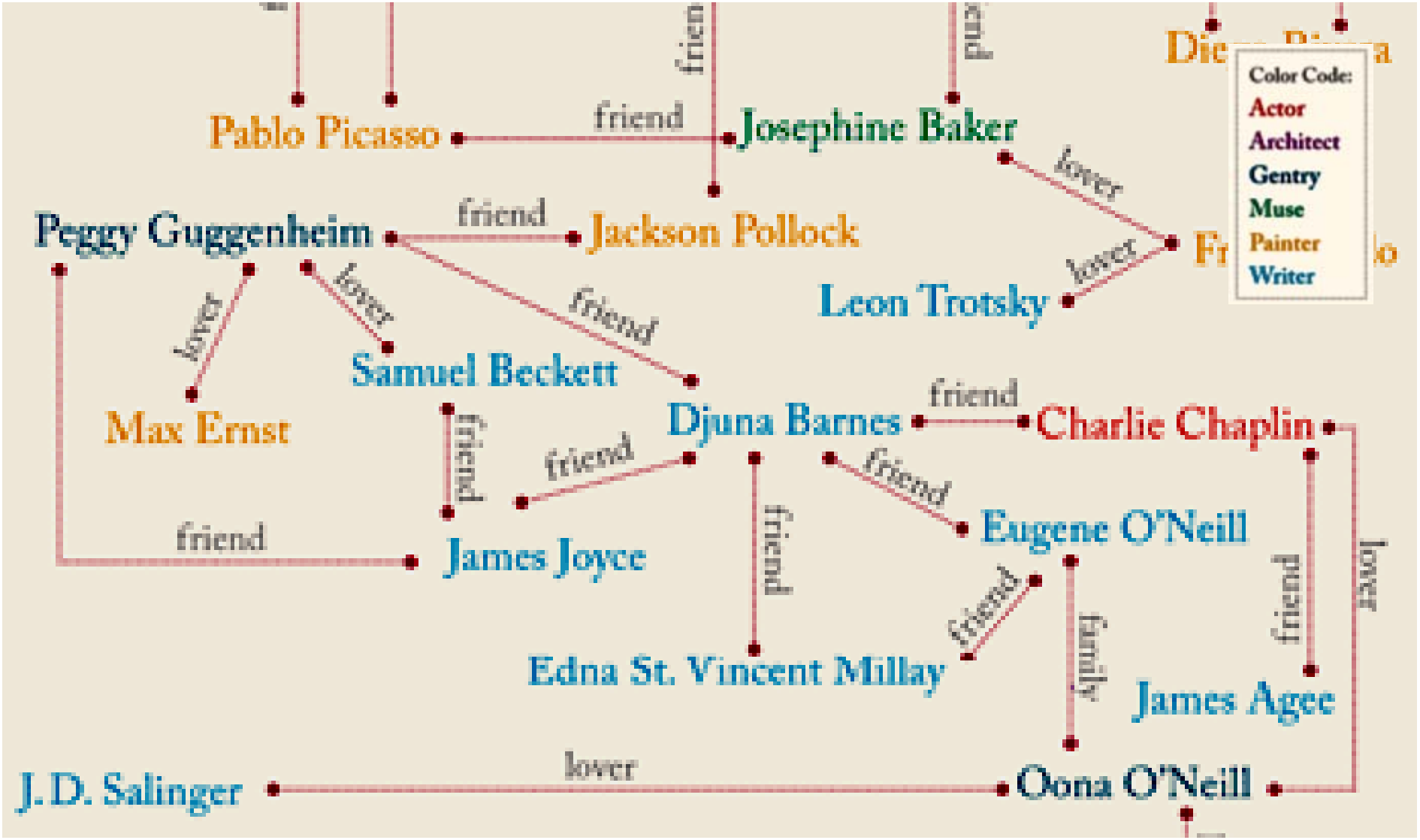


Source: http://www.mpi-fg-koeln.mpg.de/~lk/netvis/netvis/GermanyInc.html

Ego-centered Network



Social Networks of famous painters, writers, actors of XX century



Source: Friends, Lovers, and Family http://laphamsquarterly.org/visual/charts-graphs/?page=78

Centrality

Prominence

- The study of structural properties of networks and their interplay with the processes taking place on the network is one of the main problems in the last years in the field of complex network analysis
- **A primary use of graph theory in social network analysis is to identify "important" actors.**
  Centrality and prestige concepts seek to quantify graph theoretic ideas about an individual actor's prominence within a network by summarizing structural relations among the graph nodes.
- An actor's prominence reflects its greater visibility to the other network actors (an audience). An actor's prominent location takes account of the direct sociometric choices made and choices received (outdegrees and indegrees), as well as the indirect ties with other actors. The two basic prominence classes:
  - **Centrality**: Actor has high involvement in many relations, regardless of send/receive directionality (volume of activity)
  - **Prestige**: Actor receives many directed ties, but initiates few relations (popularity > extensivity)

Source: Wasserman&Faust "Social Network Analysis" (W&F)

Centrality: degree, closeness, betweenness

- Originally, the centrality (or power measure) was studied particularly in the social science from the perspective of the social interaction, but nowadays it is also considered for other different complex networks (such as web)
- The centrality measures quantify each node influence and importance among the rest of the nodes in the processes that take place in the network. The best known measures are three "classical":
  - the degree centrality
  - the closeness centrality,
  - the betweenness centrality
- and
  - the eigenvector-like centralities

Source: M. Romance, Journal of Computational and Applied Mathematics (2010)

Centrality: Eigenvector Centrality

- Eigenvector centrality was introduced by Phillip Bonacich in 1987
- “Google's workhorse search engine ranking algorithm, PageRank, is actually a variant on an SNA concept - Bonacich Power Centrality. Bonacich (1987) hypothesized that someone's power in society depends on the power of his or her social contacts. Bonacich formalized this mathematically:

  $$c_i = B(c_1R_{i1} + c_2R_{i2} + \ldots + c_nR_{in}) ,$$

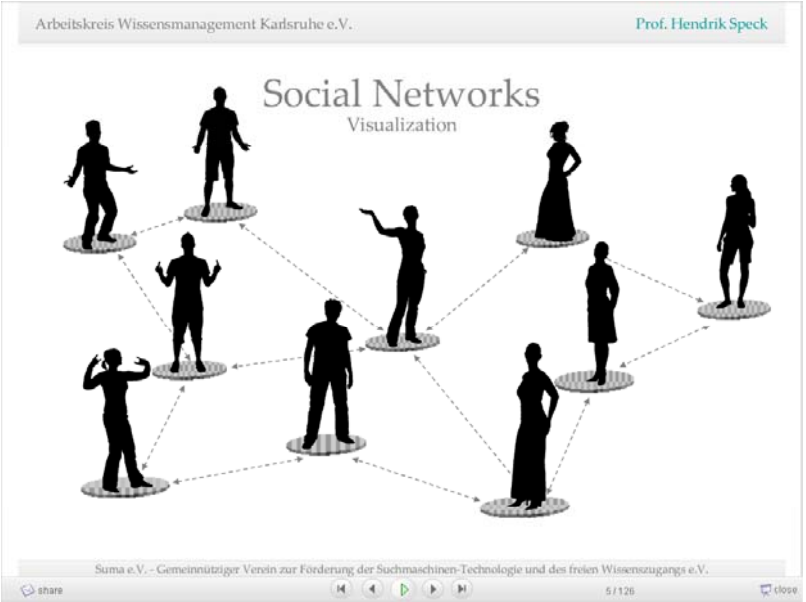
  where c_i is the person in question, B is the magnitude of the effect, and R_{ij} is the strength of the relationship between the person in question, i, and each of the other people, j, under consideration.

  If B=1 , the formula becomes eigenvector centrality, of which PageRank is a variant. Now, Page, et al. (1998) do not cite Bonacich, I am not claiming that they stole the idea - I am merely stating that a social network analyst appears to me to have been the first to think up the concept”.

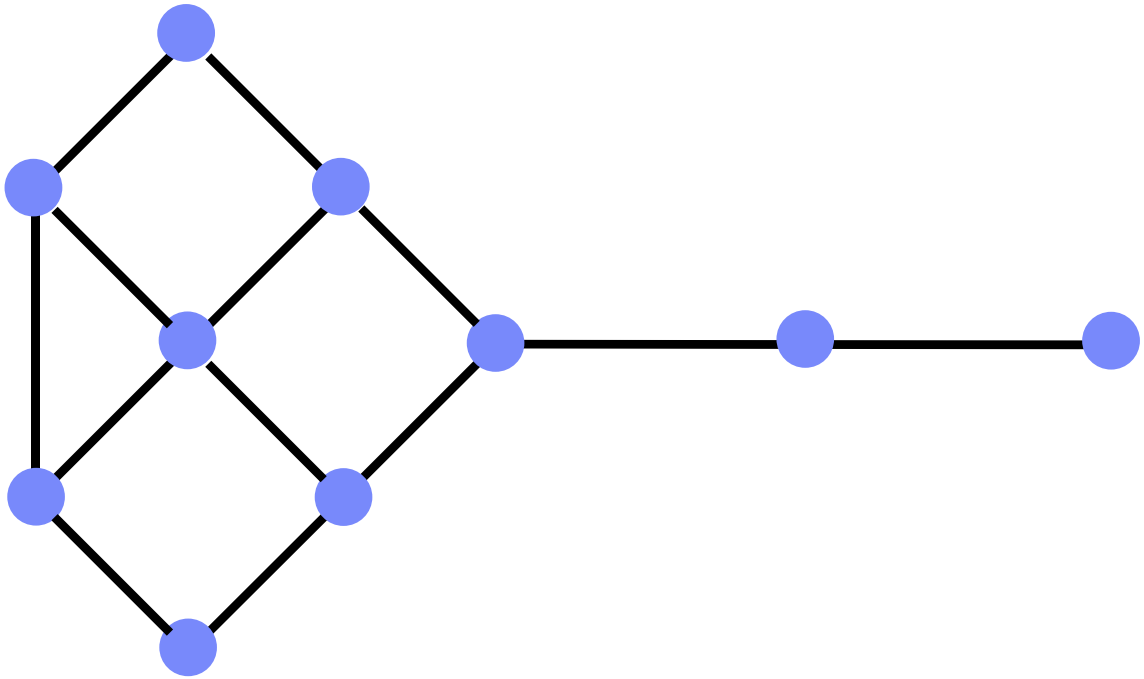
  Solomon Messing http://www.stanford.edu/~messing/RforSNA.html

Social networks

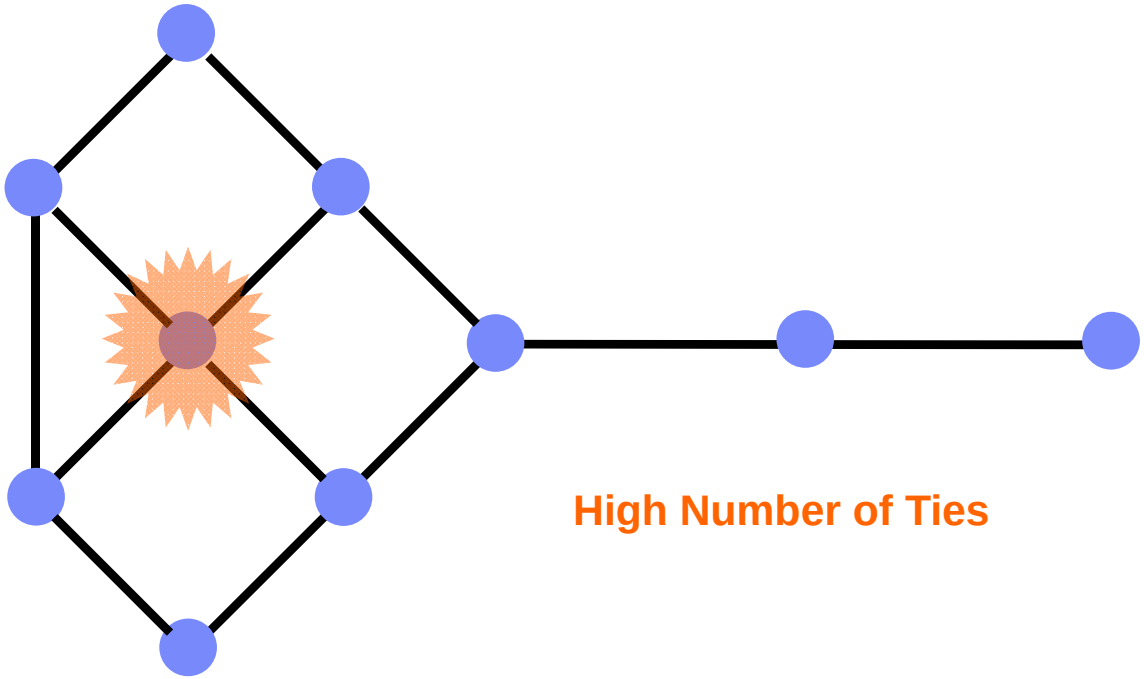
- SNA focus on relationships between actors
- “Social network analysis reflects a shift from the individualism common in the social sciences towards a structural analysis”.
  Garton et al Studying Online Social Networks



- Graphics from Prof. Hendrik Speck's tutorial at 5th Karlsruhe Symposium for Knowledge Management in Theory and Praxis, 2007

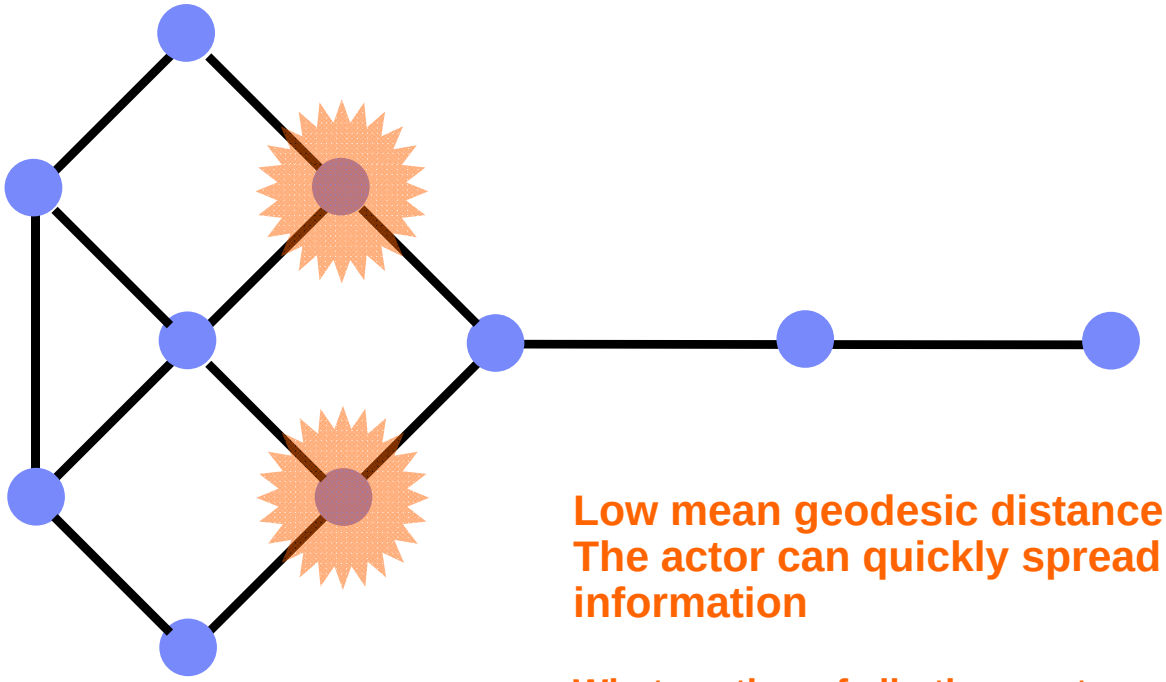


Degrees Centrality



High Number of Ties

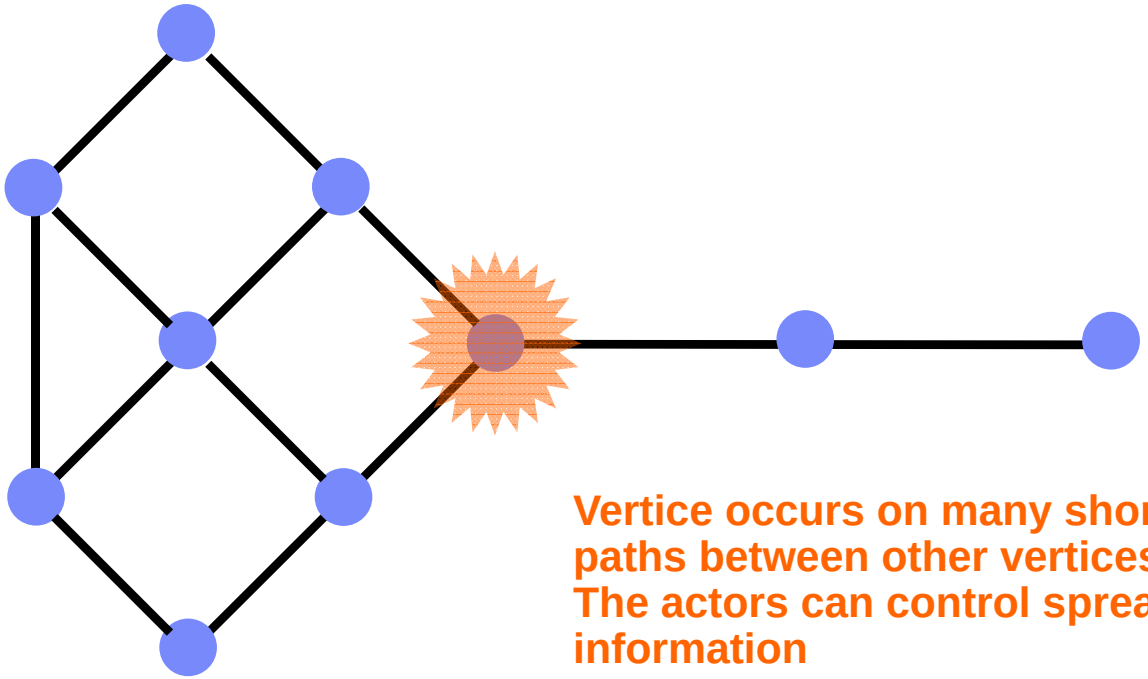
Closeness Centrality



Low mean geodesic distance
The actor can quickly spread information

What portion of all others actors can the actor reach in one step, two steps, three steps, etc.

Betweeness Centrality



Vertice occurs on many shortest paths between other vertices
The actors can control spread of information

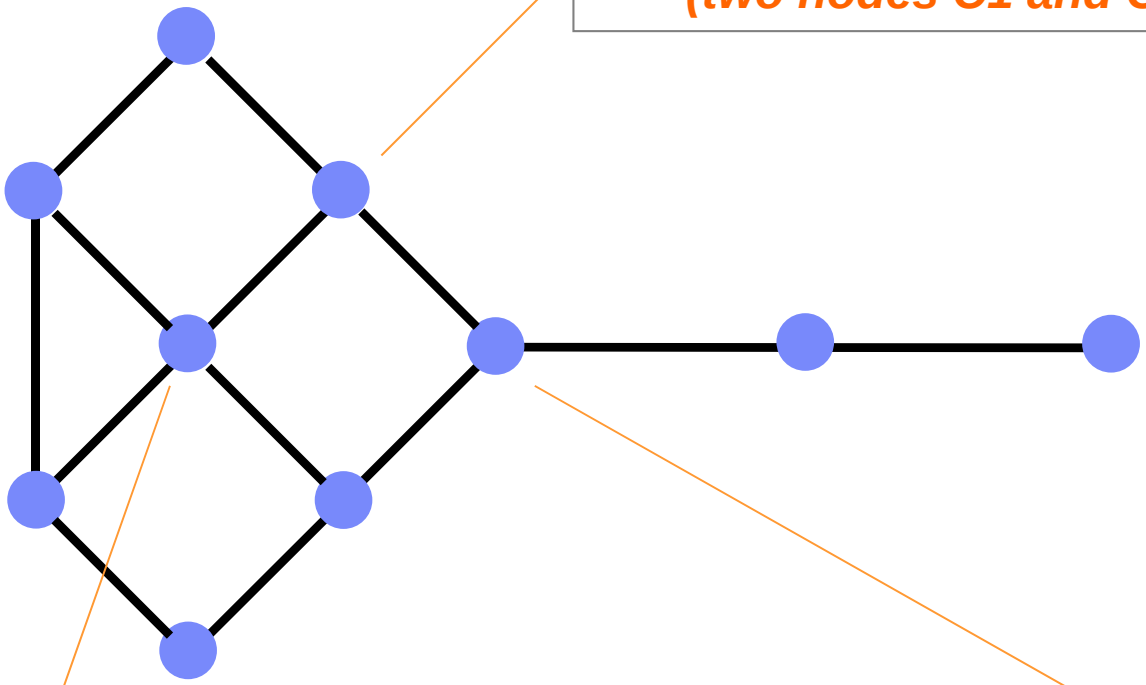
Source: Hanneman, Robert A. and Mark Riddle. 2005. Introduction to social network methods. (http://faculty.ucr.edu/~hanneman/)

Betweennes centrality...

- Betweenness centrality
  - For example, let's suppose that I wanted to try to convince the Chancellor of my university to buy me a new computer. According to the rules of our bureaucratic hierarchy, I must forward my request through my department chair, a dean, and an executive vice chancellor. Each one of these people could delay the request, or even prevent my request from getting through. This gives the people who lie "between" me and the Chancellor power with respect to me. To stretch the example just a bit more, suppose that I also have an appointment in the school of business, as well as one in the department of sociology. I might forward my request to the Chancellor by both channels. Having more than one channel makes me less dependent, and, in a sense, more powerful.

Source: Hanneman, Robert A. and Mark Riddle. 2005. Introduction to social network methods. (http://faculty.ucr.edu/~hanneman/)

Many Faces of “Power”

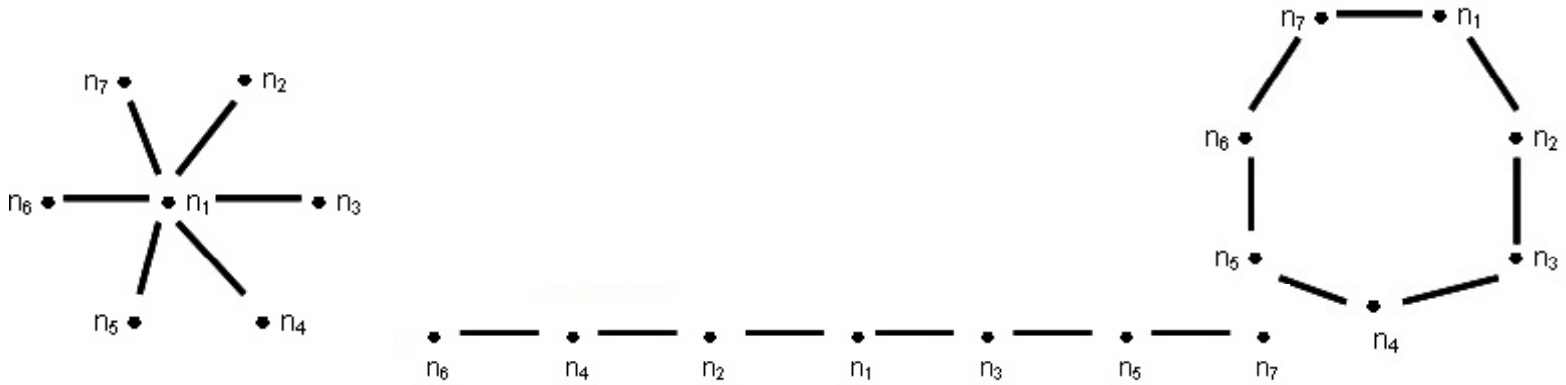


C – Closeness Centrality (two nodes C1 and C2)

D – Degree Centrality (one node)

B – Beetwenness Centrality (one node)

Another Example (W&F)

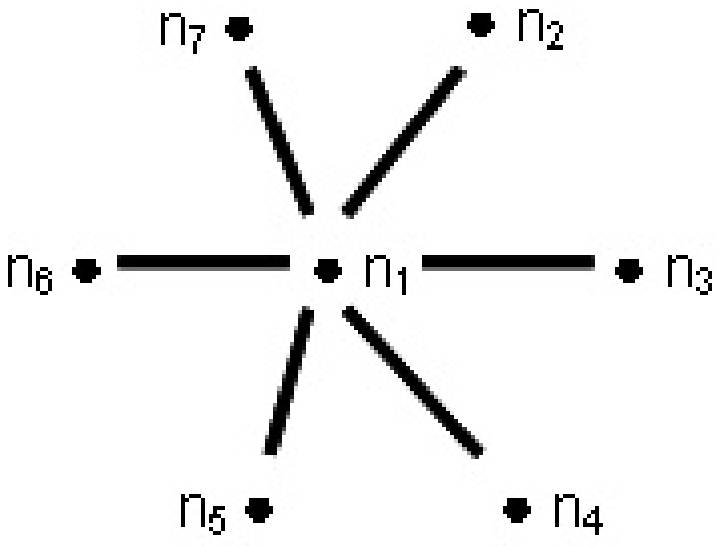


- Star Line graph (chain), Circle

Source: Wasserman&Faust "Social Network Analysis“ (W&F)

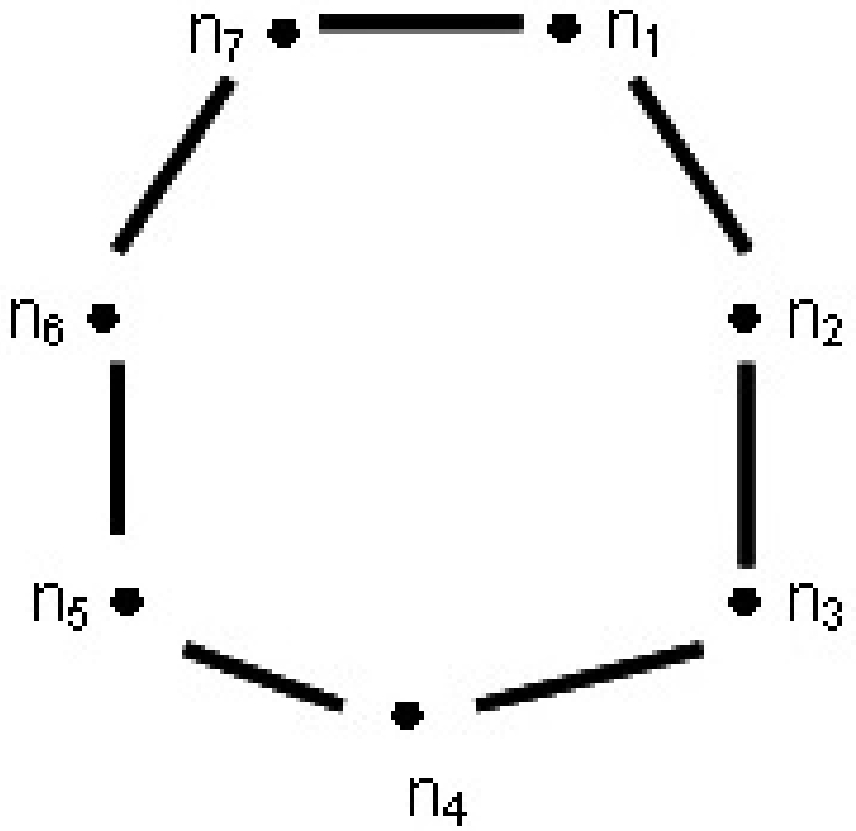
Star

- Degree Centrality STAR: the most central actor (n1) has degree centrality = 6 but the six peripheral actors each have degree centrality = 1;
- Closeness Centrality STAR: In the star graph, actor n1 has closeness = 1.0 while the six peripheral actors = 0.545.
- Betweenness Centrality STAR: In the star graph, actor n1 has betweenness = 1.0 while the six peripheral actors = 0.0.



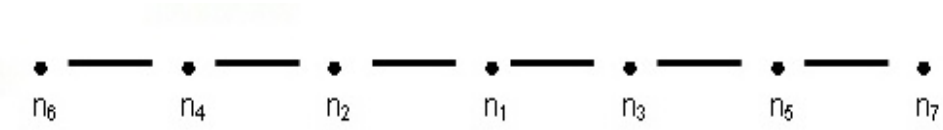
Circle

- Degree Centrality CIRCLE: All seven circle graph actors have identical degree centrality (=2), so no central actor exists; their standardized values are each 0.333.
- Closeness Centrality CIRCLE: All circle graph actors have the same closeness (0.50).
- Betweenness Centrality CIRCLE: All circle graph actors have the same betweenness (0.2).



Chain

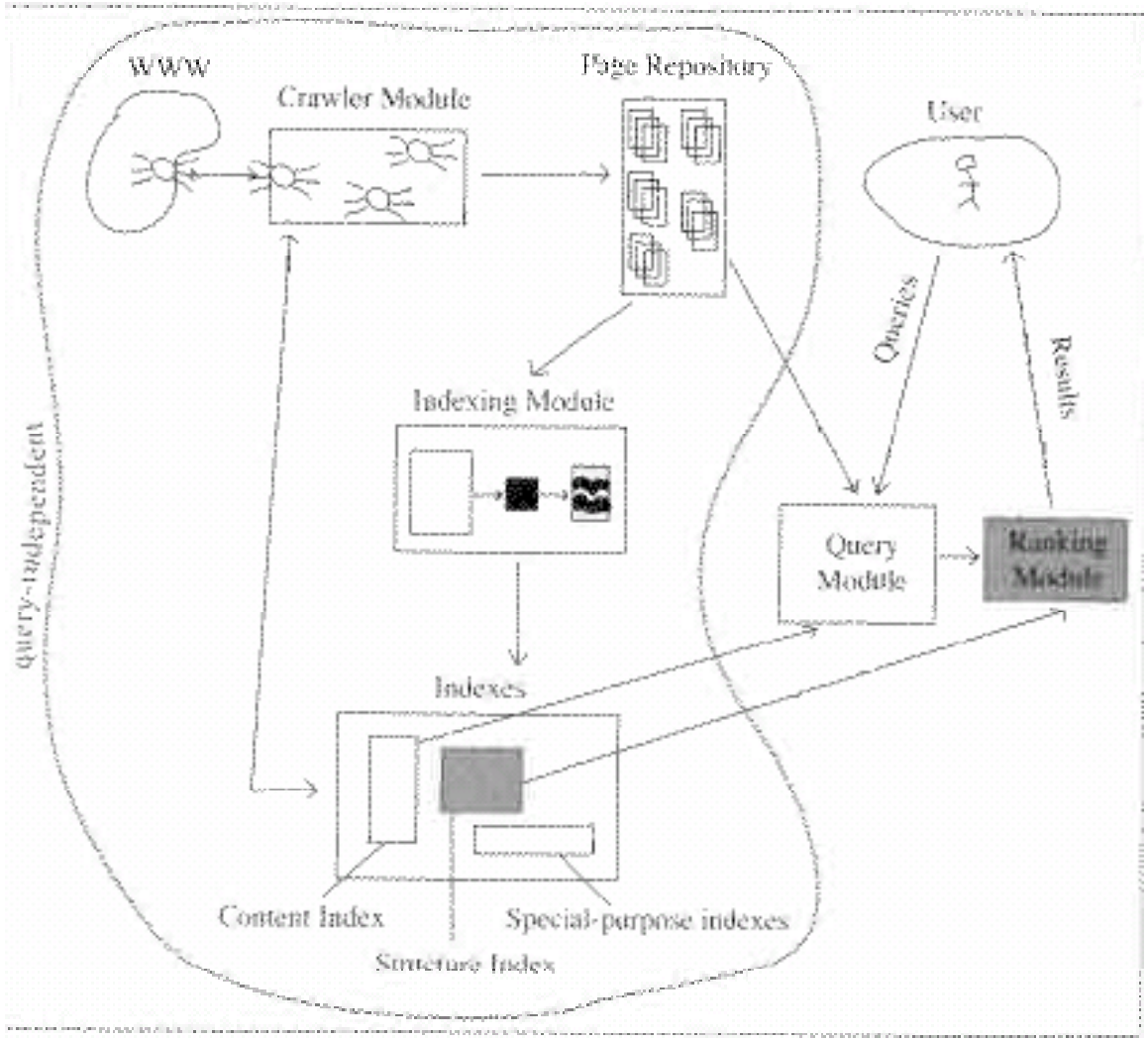
- Degree Centrality LINE: In the line graph, the two end actors have smaller degree centralities (degrees = 1) than those in the middle (=2); the respective standardized scores are 0.167 and 0.333.
- Closeness Centrality CHAIN: In the chain graph, the two end actors are less close (0.286) than those in the middle (0.50).
- Betweenness Centrality CHAIN: In the chain graph, the two end actors have no betweenness (0.0), the exactly middle actor n1 has the highest betweenness (0.60), while the two adjacent to it are only slightly less central (0.53).



PageRank

Sergey Brin (right) and Larry Page (left), 1998

Elements of a search engine



PageRank

According to Google,

- “PageRank relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at more than the sheer volume of votes, or links a page receives; it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important".

- A page that is linked to by many pages with high PageRank receives a high rank itself.

PageRank

- The value of the page link-votes is divided among all the outbound links on a page.
- PageRank PR**(*)** conferred by an outbound link *L***(*)** is equal to the document's own PageRank score divided by the normalized number of outbound links (it is assumed that links to specific URLs only count once per document).

$$PR(A) = PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D) + \ldots$$

- i.e. the PageRank value for a page A is dependent on the PageRank values for each page *v* out of the set $\mathbf{S_A}$ (this set contains all pages linking to page A), divided by the number $L(v)$ of links from page *v*.

PageRank with a damping factor

$$PR(A) = (1-d)/N +$$

$$d[PR(B)/L(B) + PR(C)/L(C) + PR(D)/L(D) + \ldots]$$

Usually, *d* is set about 0.85, *N* is the total number of pages

Computation of the PageRank vector

- Solve the following eigenvector problem:
  - $r^T = r^T G$
  - $r^T e = 1,$

where r(A) =PR(A), and G is the above Brin-Page *N*x*N* matrix

$G = dS + (1-d)/N\ ee^T$, $e^T = (11,\ldots,1)$

Alexander Troussov, Ph.D., IBM Dublin Software Lab

4th Russian Summer School in Information Retrieval, September 13-18, 2010, Voronezh

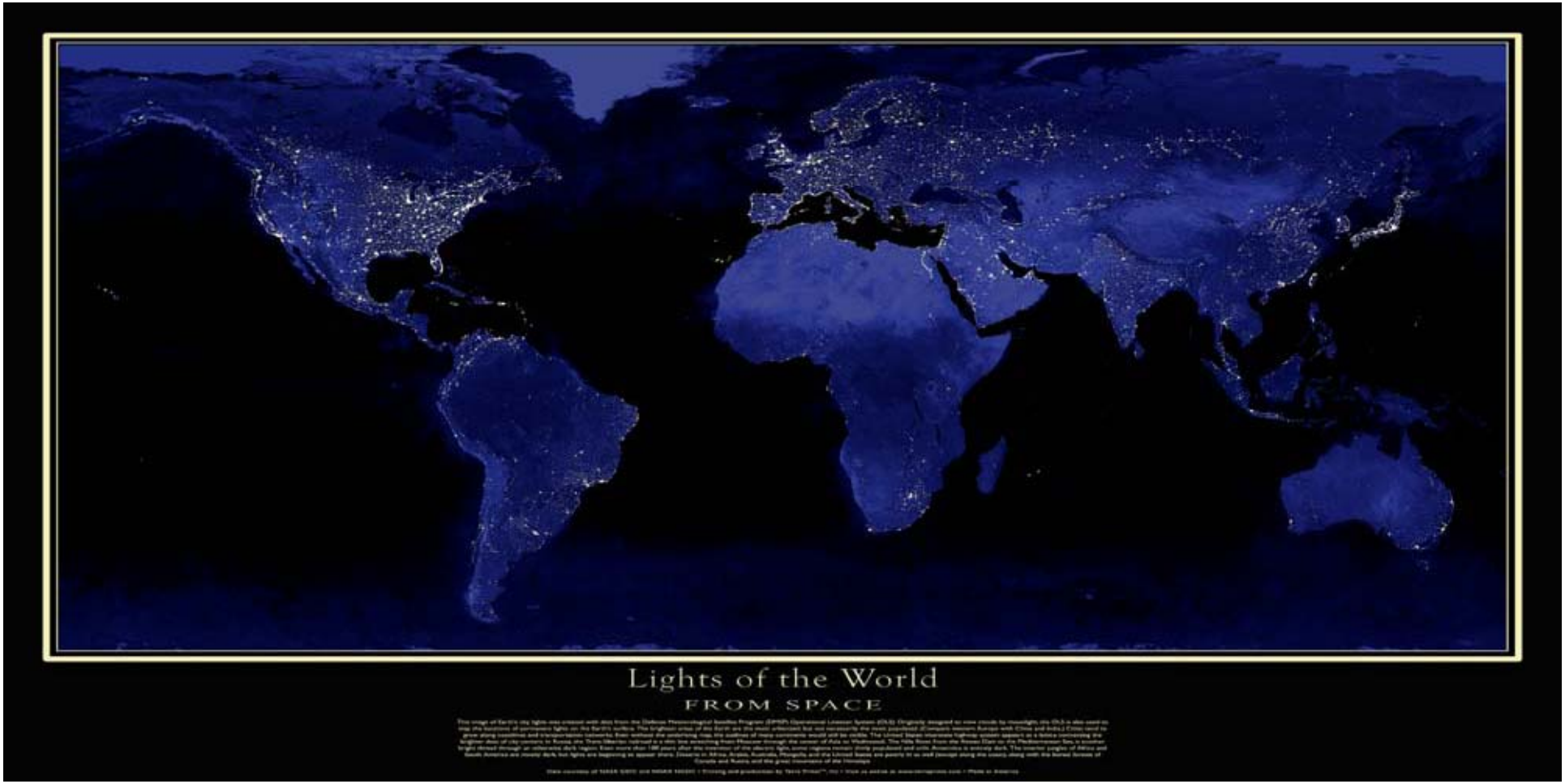
RuSSIR
Russian Summer School
in Information Retrieval

Global vs. Local Ranking and Relevancy Propagation

Ranking

- Link analysis is frequently employed for ranking and navigation
- Graph-based recommender systems should recommend
  "Important" objects (nodes, links, subgraphs)
  which are also located
  Close enough to the initial points of interests (query, initial seed)

Global Ranking (like Google's PageRank) – "Copernican" approach, a view on the network from external point



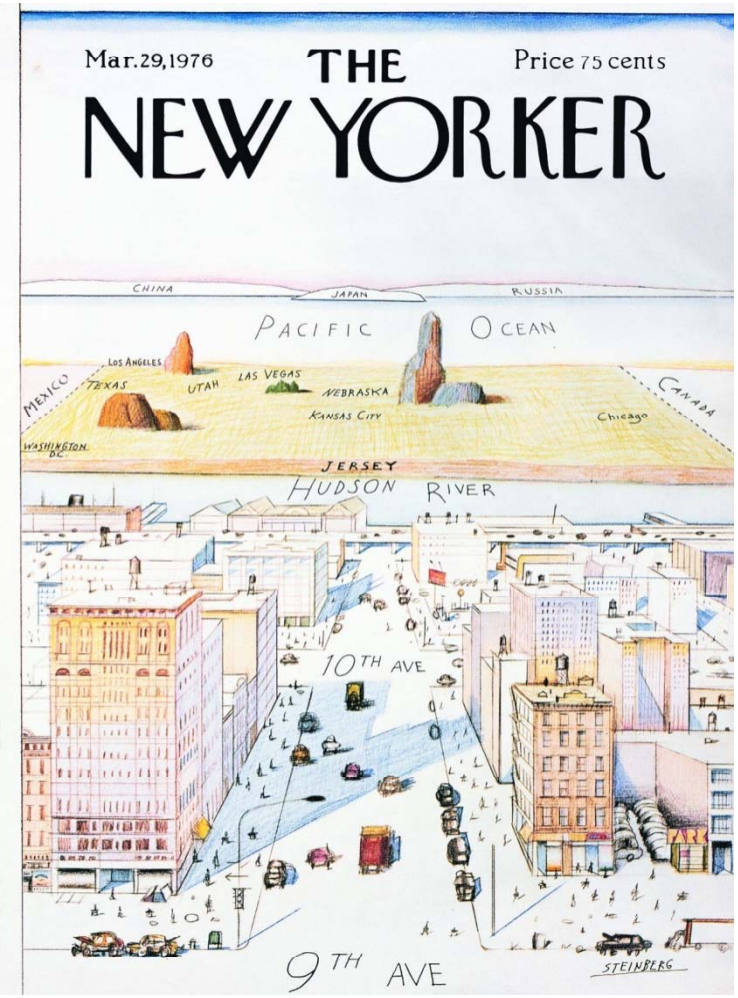
Source: NOAA

From Global Ranking to Local Ranking

- Graph-based recommender systems should recommend
  "Important" objects (nodes)
  which are also located
  Close to the initial points of interests (query, initial seed)
- One of the leading approaches in recommenders is:
  Results of Global Ranking (Link analysis)
  are "filtered" according to their proximity to the query
- Why not "Local Ranking" which takes into account the proximity to the query?

Source: ….

Local Ranking – is needed for recommenders should rely on Ptolemaic view



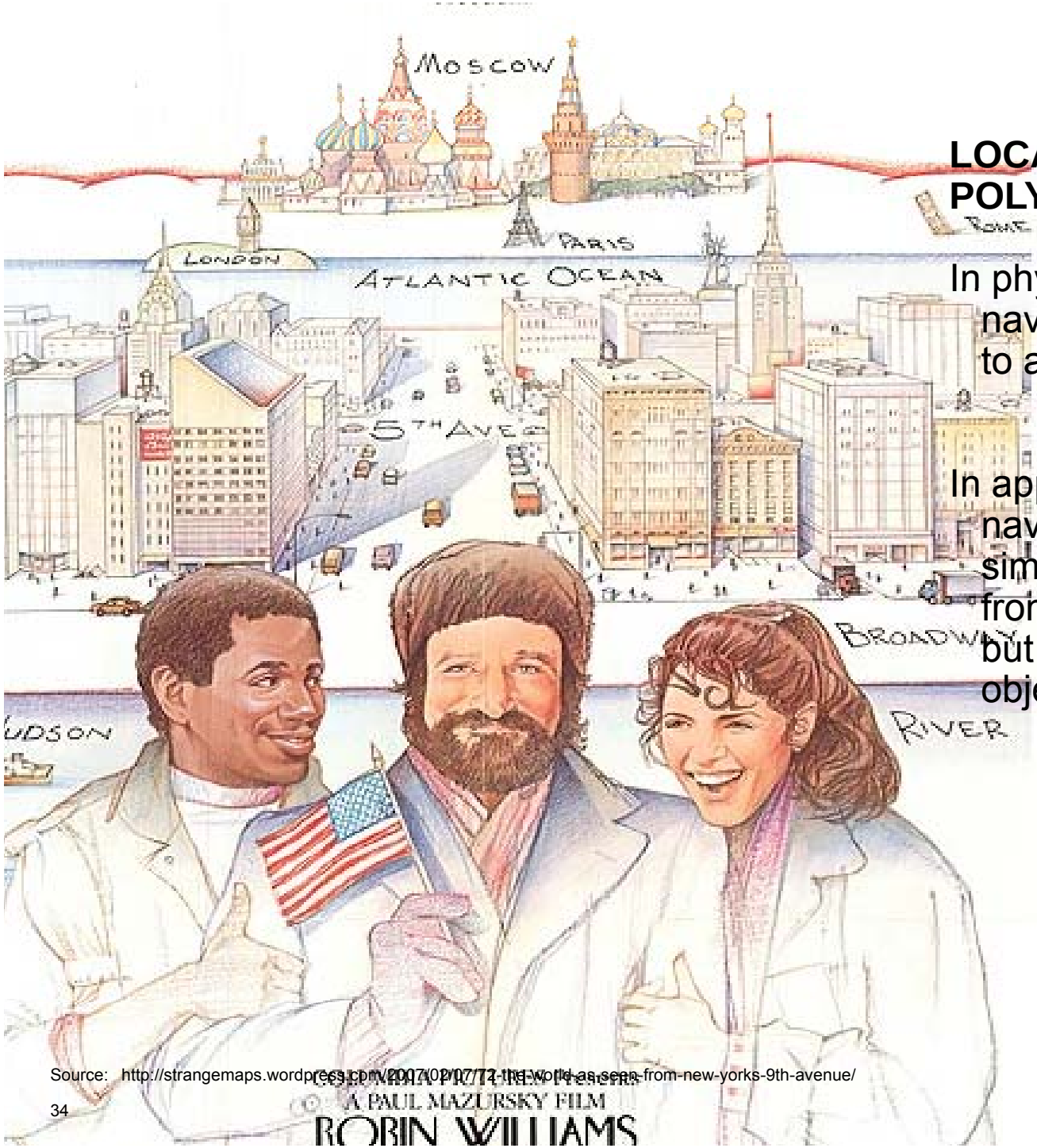
LOCAL RANKING: EGO-CENTERED

Ego-centered or "personal“ networks provide a Ptolemaic views of their networks from the perspective of the persons (egos) at the centers of their network.

Such view is consistent with navigation in the physical universe in which we live.

Such view is consistent with the perception of an actor in a social network.

Source: http://strangemaps.wordpress.com/2007/02/07/72-the-world-as-seen-from-new-yorks-9th-avenue/



LOCAL RANKING: POLY-CENTERED

In physical space – navigation is from one point to another.

In applications to virtual spaces navigation is not simply browsing from a single object to another, but by dealing with several objects at the same time .

Source: http://strangemaps.wordpress.com/2007/02/07/72-the-world-as-seen-from-new-yorks-9th-avenue/

Tasks

Link Prediction

- Should we introduce Mr J. to Mr T.?
- Closeness, proximity, relevance, similarity

An example of egocentric queries

– Social spaces such as blogs, wikis and online social networking sites are enabling the formation of online communities where people are linked to each other through direct profile connections and also through the content items that they are creating, sharing, tagging, etc. The Semantic Web provides a platform for gathering diverse information from heterogeneous sources and aggregating such linked data into multidimensional network of nodes representing people, organisations, projects etc. Spreading activation methods were used in [Kinsella et al., 2008] to augment objects from social spaces, by highlighting related objects, recommending tags, and suggesting relevant sources of knowledge.

Source: Kinsella et al. Navigating and Annotating Semantically-Enabled Networks of People and Associated Objects. 2008

SA based recommender suggests to connect with Tim

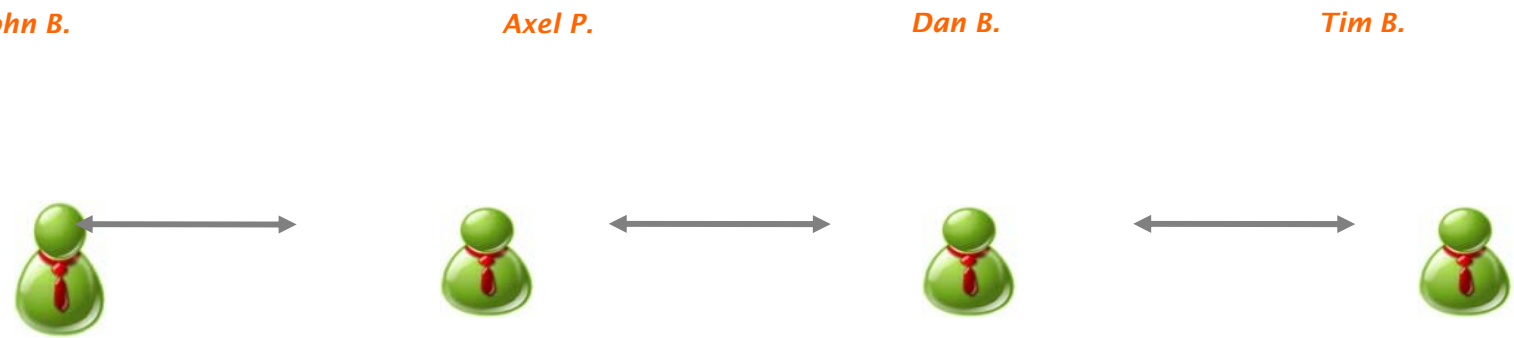
John B.

?

Tim B.

38

Tim is three steps away from John



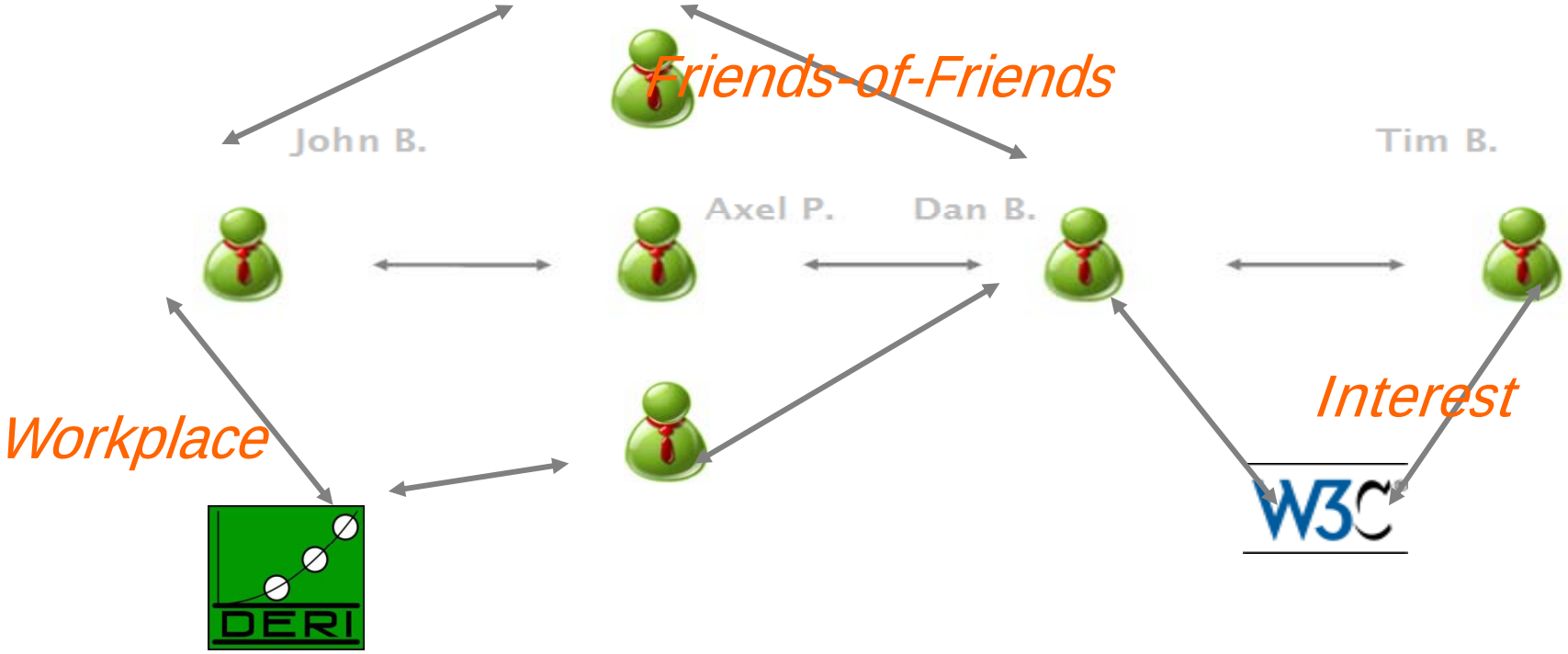
“Three steps a

Why recommender decided that this three steps away connection is a strong connection?

40

John and Tim –

Recommender computes that this is a strong connection because of multiple ways of connections



John and Derek

Recommender computes that such type of connectivity is a weak connection

Spreading Activation for ambient navigation

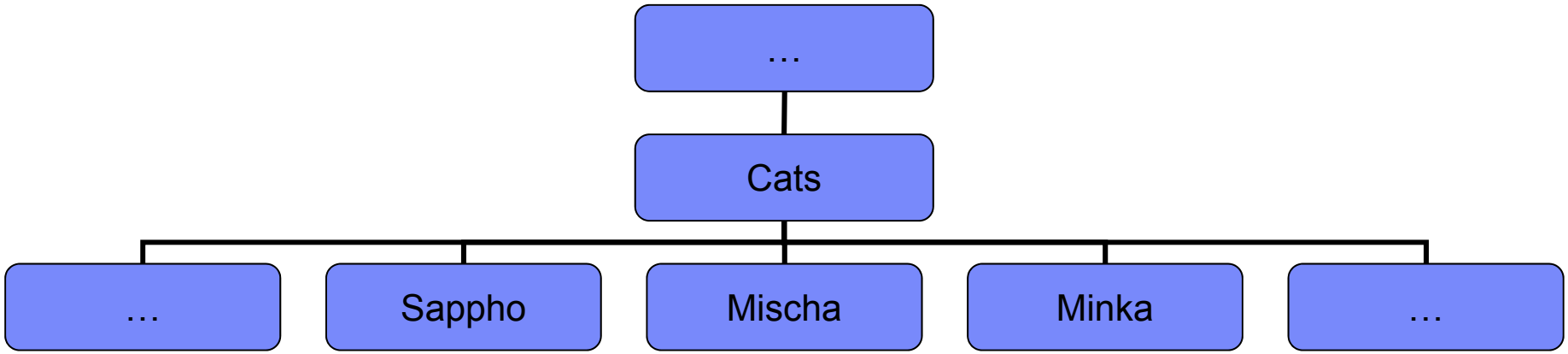
- Spreading activation engine (IBM) was used in EU Project Nepomuk
  - No SPARQL or other queries
  - No browsing
  - Fast
  - Shows something of cognitive interest
  - to perceive, contextualize, simplify, and make sense of otherwise complex interlinked data
  - without cognitive load: How can I ask for what?

Using SA for ego-centric queries

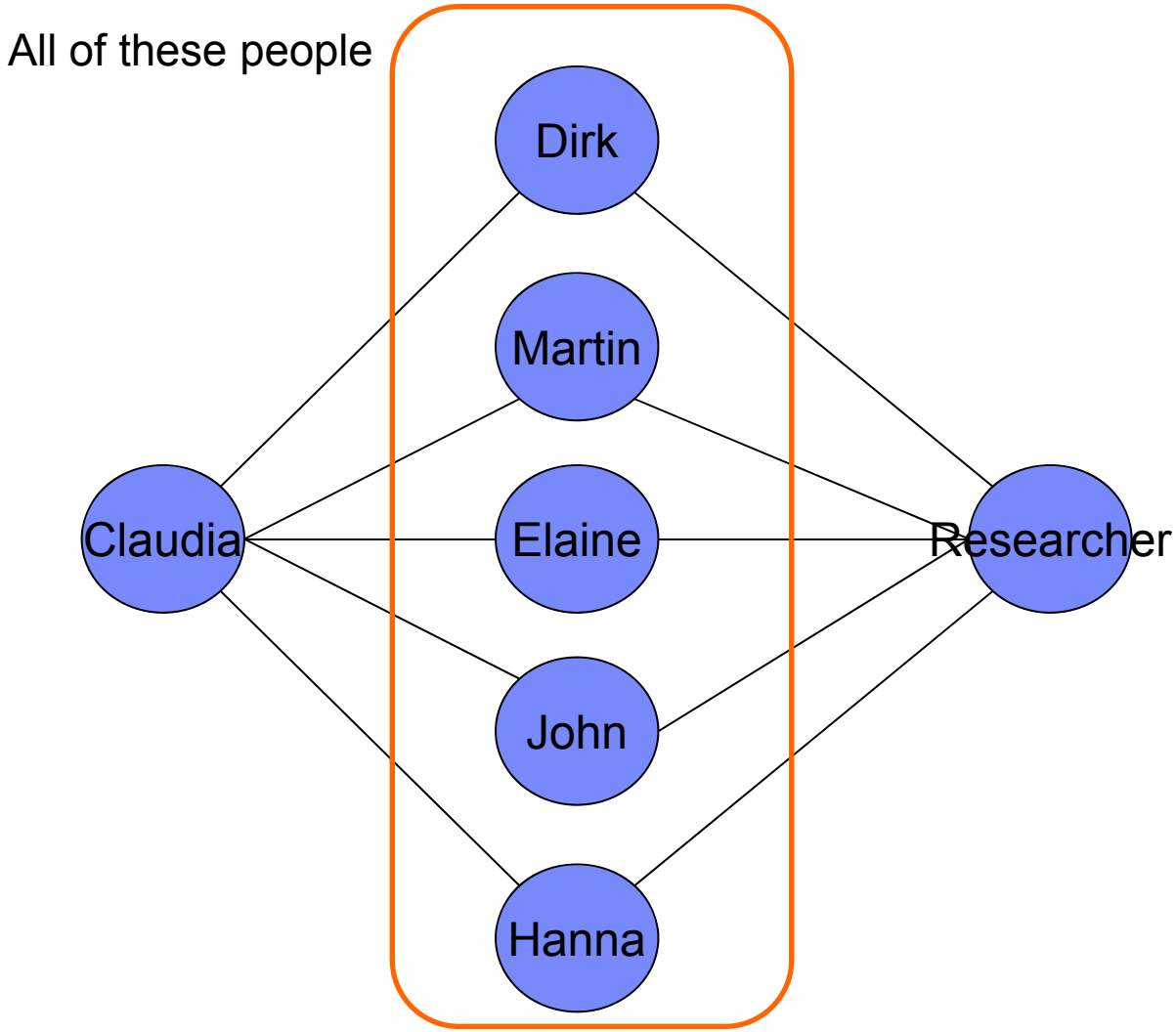
- On the next slides we show how SA performs for egocentric queries
  - finds another instances of the same class, provides generalisation, finds common features of “neighbours”
  - and ranks these findings in a sensible way, depending on the topology of the cognitive map

    - With whom is *Claudia* connected?

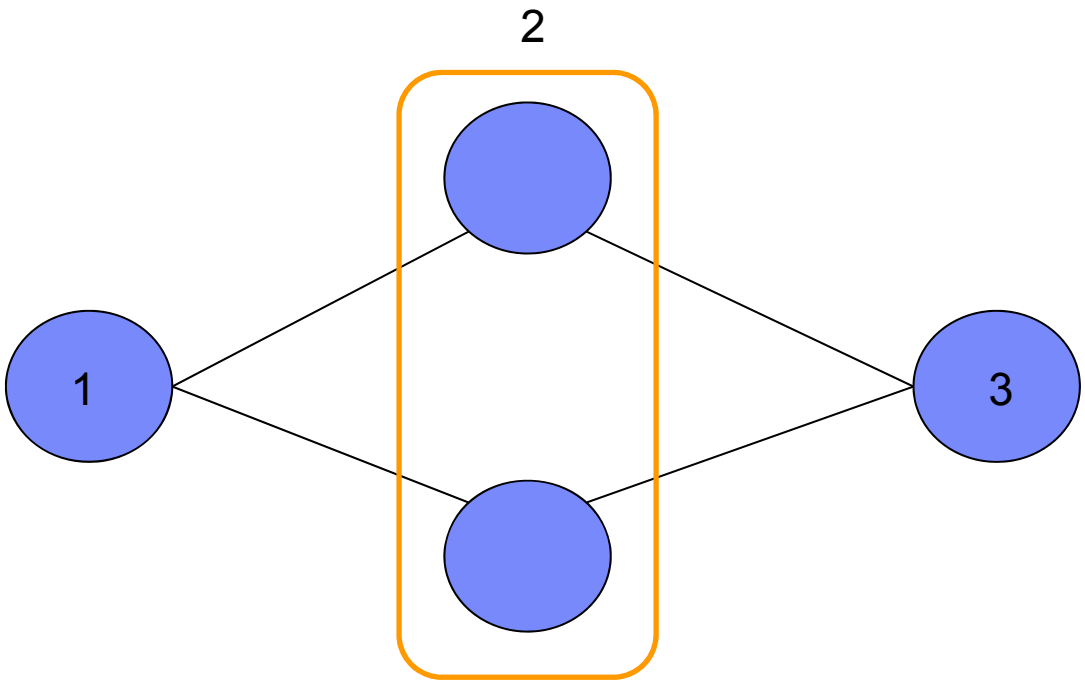
Generalisation



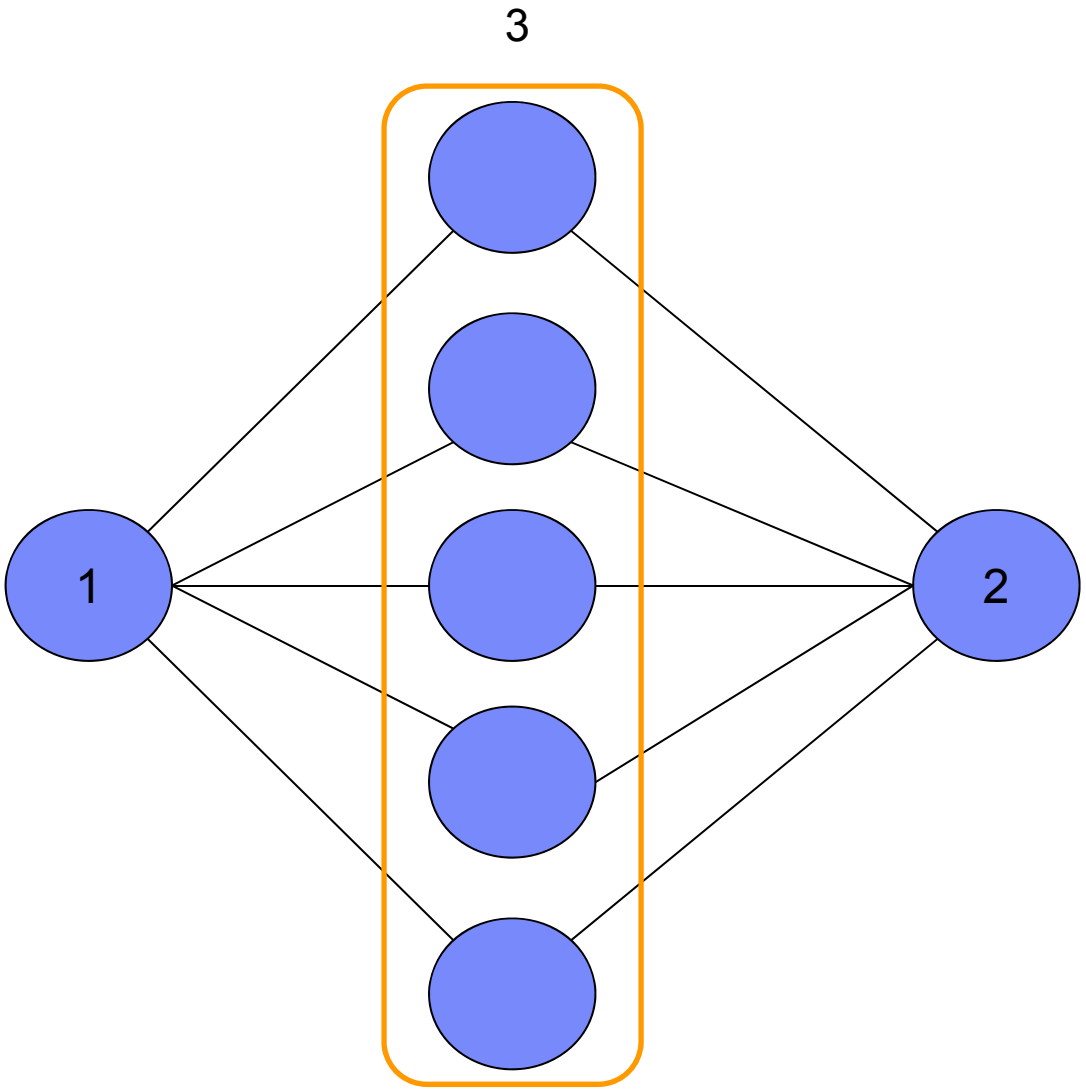
With whom is Claudia connected?



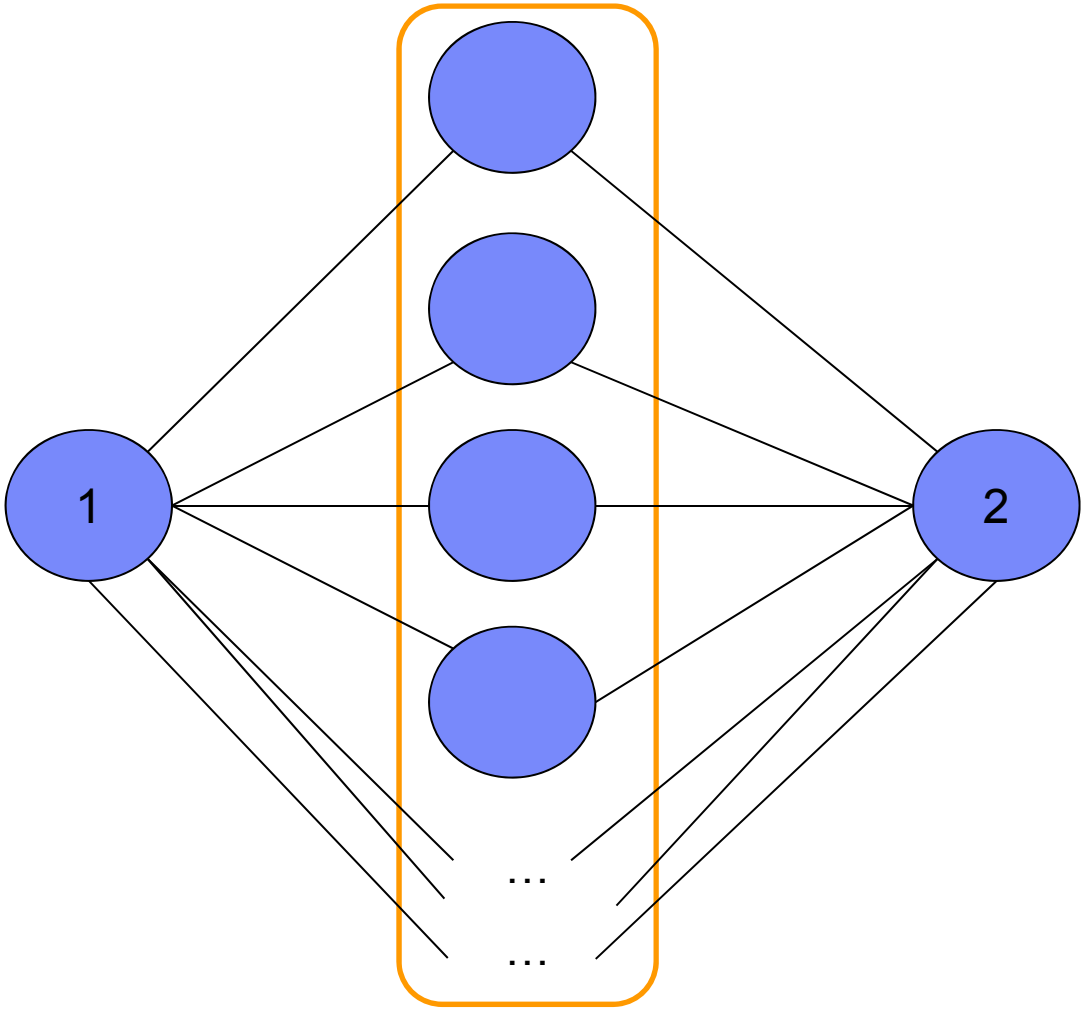
Ranking



Ranking



Ranking



Ambient Navigation – Theoretical view

- "Ambient Navigation" is our user-centric generalisation of "dynamic taxonomies"
  - multidimensional networks (like PIMO) provide a single, coherent framework in which users can focus on one or more nodes (concepts) in the network, and immediately see a conceptual summary of their focus,
    - in the form of a reduced network derived from the original one by pruning unrelated concepts
    - augmented with relations "strong connectivity"
  - Concepts in the transformed network can be used to set additional, dependent foci and users iterate in a guided yet unconstrained way until they reach a result set sufficiently small for manual inspection

Many-to-many navigation (ambient navigation)

- Nepomuk Simple: a Pile-based UI, which is under development by KTH for the EU project Nepomuk.
- We conclude that the process of incremental modification of a “pile”, assisted by recommendations resulting from fuzzy polycentric queries, might be formally viewed as the process of browsing and exploring data provided by networks.
  - This type of navigation is not simply browsing from a single object to another, but by dealing with several objects at the same time in a process similar to how one browses in a library or shop. Such an “ambient navigation” tool might be used for exploring various massive multidimensional networks, which occur in the socio-semantic information space which we encounter in the modern information age. This usage extends beyond the scope of scenarios explored in the EU 6th framework project Nepomuk where these ideas have been developed. For instance, Nepomuk Simple powered by Galaxy might be used to navigate not Nepomuk PIMO, but social networks.

RuSSIR
Russian Summer School in Information Retrieval

Web and Communities

Community detection … but What is Community?

- Are you Russian? Yes. Are you Irish? Yes, Are you mathematician? Yes. Are you practitioner? Yes.
  - Communities easily overlap, multiple membership and fuzzy belongings
- Remember "Strange Case of **Dr Jekyll and Mr Hyde**" (Robert Louis Stevenson, 1886).
  - How Google had failed to understand an essential property of real-world social networks
  - So by testing their social service inside a single context (Google employees only), the developers failed to notice that in real life, **people participate in multiple contexts (family, work, friends, etc) that they work actively to keep separate.** The reasons for wanting to keep these groups separate can range from wanting to keep an illicit affair secret from your spouse to political activists in oppressive regimes wanting to keep certain connections secret from the government. Another important reason to keep our communities separate, is that we often play different roles - and communicate differently http://www.iq.harvard.edu/blog/netgov/2010/03/worlds_colliding.html
  - Commununities should be kept separate

Social Sciences – Communities, Networks, Groups

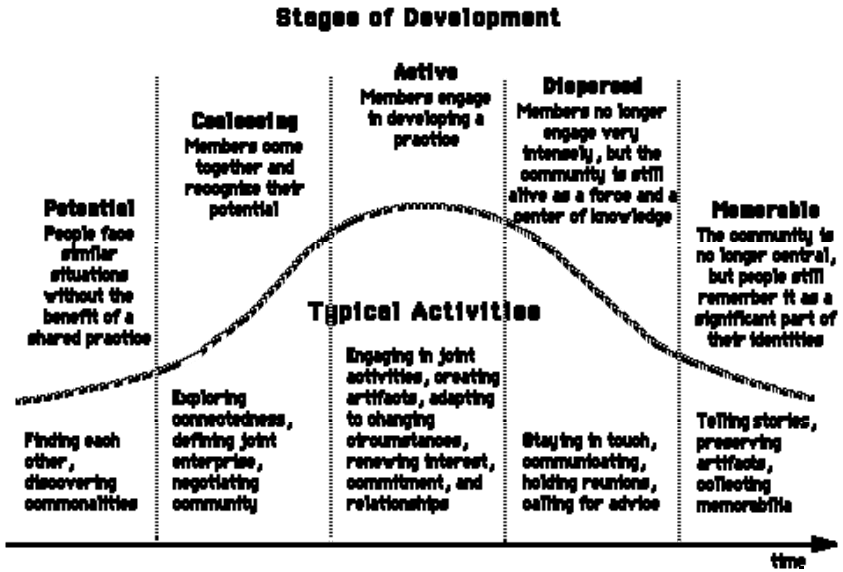
- Communities
  - In biological terms, a community is a group of interacting organisms (or different species) sharing an environment.
  - In human communities, intent, belief, resources, preferences, needs, risks, and a number of other conditions may be present and common, affecting the identity of the participants and their degree of cohesiveness.
    - Traditionally a "community" has been defined as a group of interacting people living in a common location.
    - In sociology, the concept of community has led to significant debate, and sociologists are yet to reach agreement on a definition of the term. There were ninety-four discrete definitions of the term by the mid-1950s: Community of action, Community of circumstance, Community of inquiry, Community of interest, Community of place, Community of practice
  - CoPs
- Networks – like networks of knowledge
- Groups - In social network analysis, a group is an empirically discovered structure.

Communities in Social Sciences

- Communities: A tribe learning to survive, a group of engineers working on similar problems, …
- It could take years to form a community, and community could be functional for years, decades, or much longer
- In social sciences community development is usually viewed as a lifecycle process where community moves through various stages of development characterized by different levels of interaction among the members and different kinds of activities :

Etienne Wenger (1998) Communities of Practice

The term “community of practice” is often casually used to mean everything from those working in the same workgroup to those with the same occupation to those with a common interest.



Communities: Computer Sciences

- Frequently – any empirically found group of people

- To some a virtual community of practice is a misnomer as the original concept of a community of practice (CoP) was based around situated learning in a co-located setting. However, with increasing globalization and the continued growth of the Internet many now claim that virtual CoPs do exist. For example, some claim that a wiki (such as wikipedia.org) is a virtual CoP, others argue that the essence of a community is that it is place based - a community of place. Wikipedia

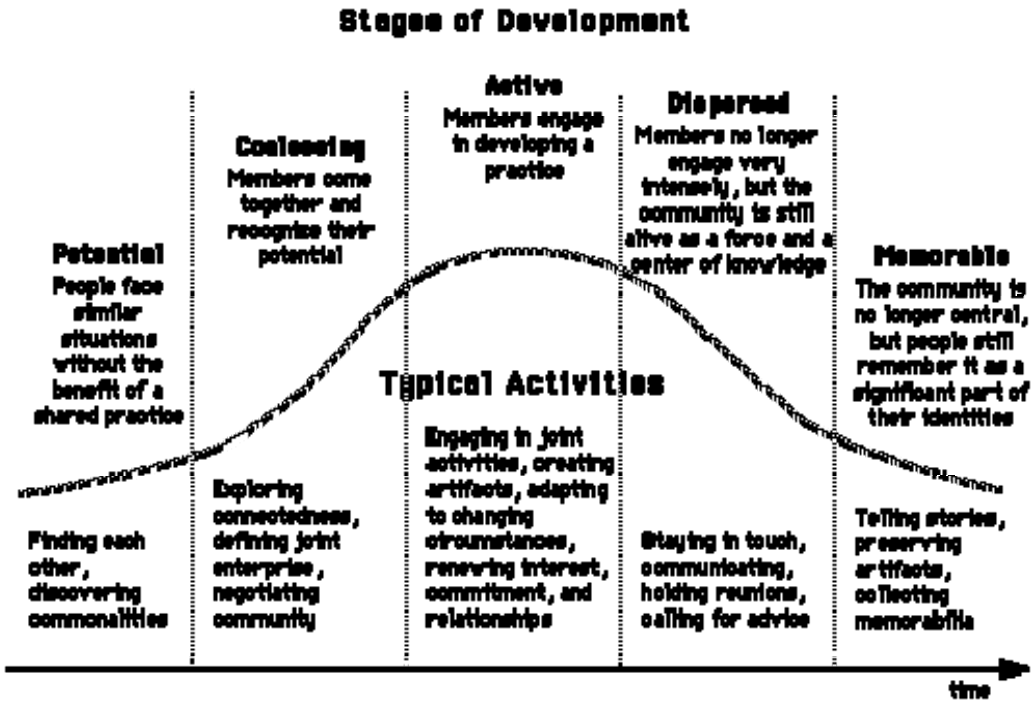
150 - Dunbar's Number

Dunbar's number, 150, describes the theoretical maximum of a genuine social network. The number is based on the limits of human abilities to identify members, relationships, and expressions

- Dunbar's number is a theoretical cognitive limit to the number of people with whom one can maintain stable social relationships. These are relationships in which an individual knows who each person is, and how each person relates to every other person.[1] Proponents assert that numbers larger than this generally require more restrictive rules, laws, and enforced norms to maintain a stable, cohesive group. No precise value has been proposed for Dunbar's number. It lies between 100 and 230, but a commonly detected value is 150.
- Dunbar's surveys of village and tribe sizes also appeared to approximate this predicted value, including 150 as the estimated size of a neolithic farming village; 150 as the splitting point of Hutterite settlements; 200 as the upper bound on the number of academics in a discipline's sub-specialization; 150 as the basic unit size of professional armies in Roman antiquity and in modern times since the 16th century; and notions of appropriate company size.
- Dunbar's number has since become of interest in anthropology, evolutionary psychology, statistics, and business management. For example, developers of social software are interested in it, as they need to know the size of social networks their software needs to take into account; and in the modern military, operational psychologists seek such data to support or refute policies related to maintaining or improving unit cohesion and morale.

Do techno-social communities have a “normal” lifecycle?

- Communities: A tribe learning to survive, a group of engineers working on similar problems, …
- It could take years to form a community, and community could be functional for years, decades, or much longer
- In social sciences community development is usually viewed as a lifecycle process where community moves through various stages of development characterized by different levels of interaction among the members and different kinds of activities :

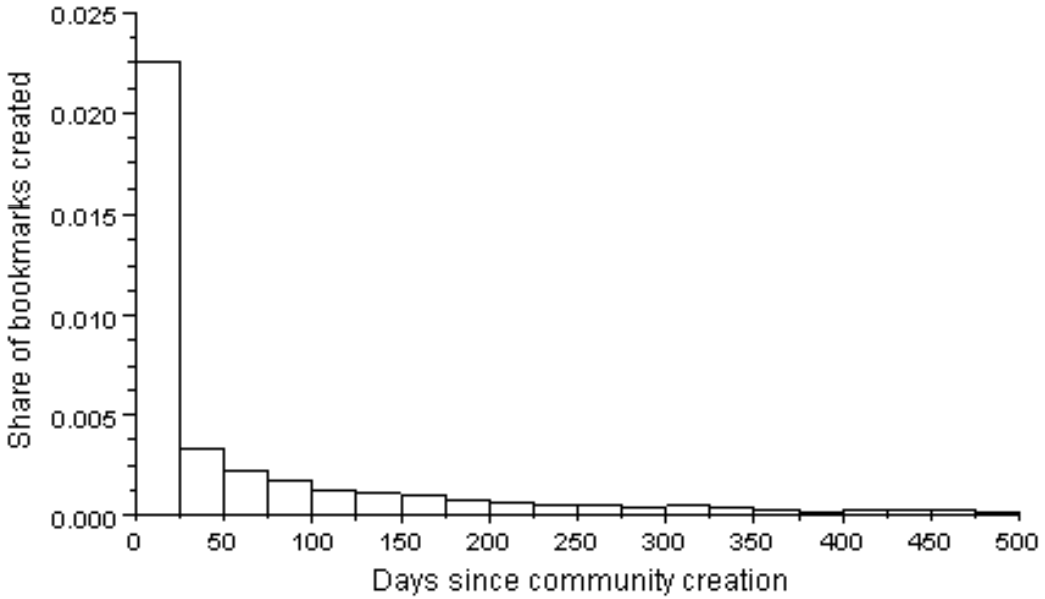


Etienne Wenger (1998) Communities of Practice

Is there an average lifespan?

- There is no average. But most communities “live fast, die young”

“Recent advances in digital technologies invite consideration of organizing as a process that is accomplished by global, flexible, adaptive, and ad hoc networks that can be created, maintained, dissolved, and reconstituted with remarkable alacrity. A central challenge, spurred by these developments, is that the nature of teams and how they are assembled has changed radically.”
Prof. Noshir Contractor

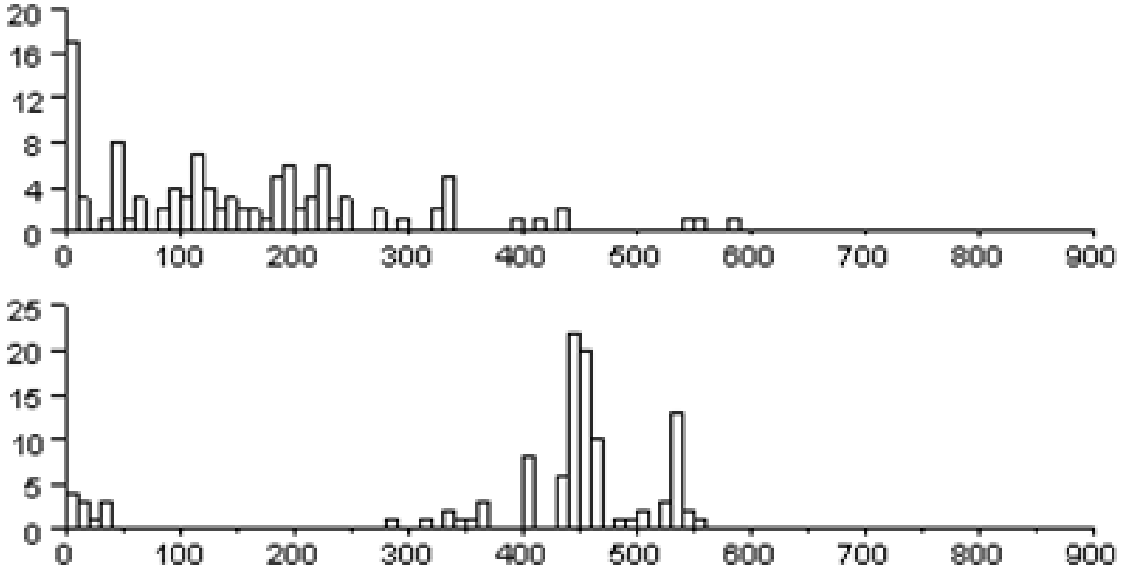


The share of bookmarks created from the moment of community formation. Possible Interpretation: on average, community members are actively involved during the first month from the moment of community creation.

Source: Nevidomsky and Troussov "Structure and Dynamics of Enterprise 2.0 Communities". WebSci 2010

What about the lifecycle?

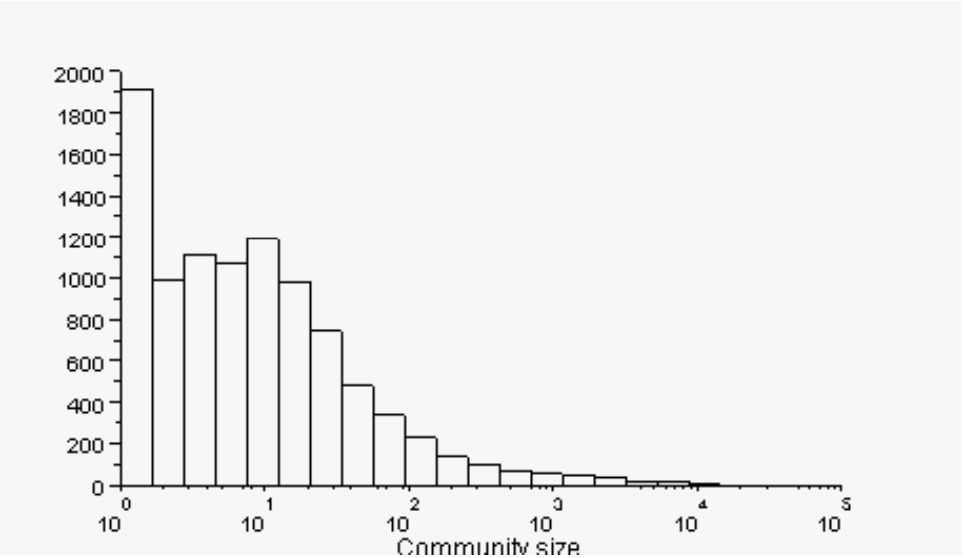
- Many communities live fast, but some definitely show long and fruitful life.



Number of bookmarks created in two communities about 100 members each, within the first 900 days.

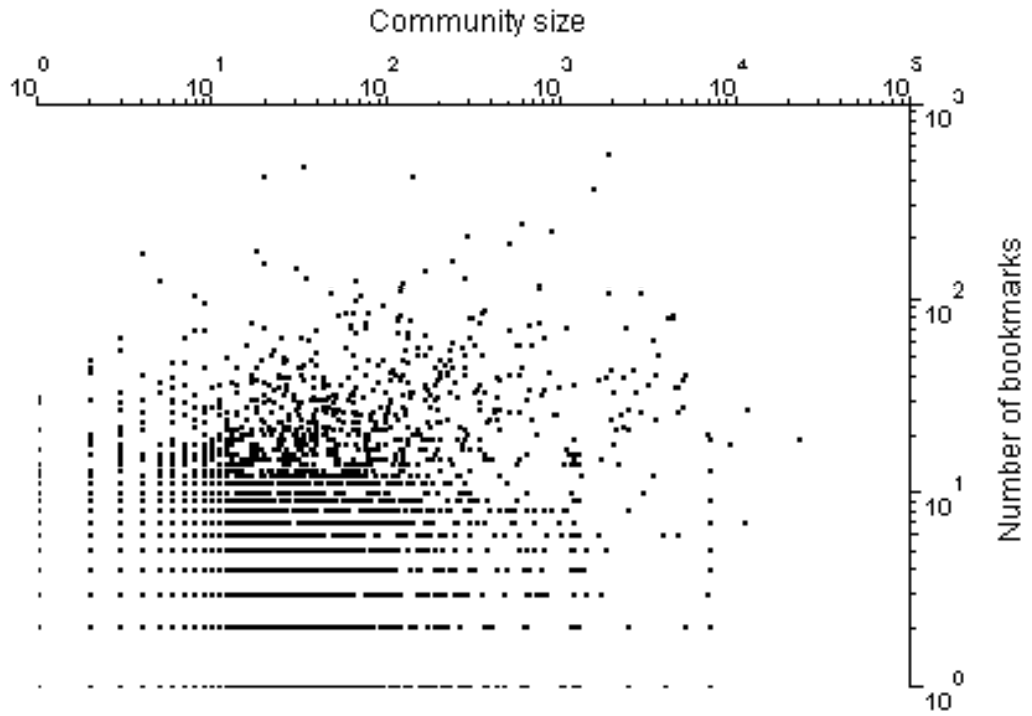
What is an average Big Blue community?

- Communities formed in IBM in Communities application of IBM Lotus Connections suite:
  - More than 9500 Communities,
    The oldest – 3 years,
    The biggest – 23,000 members
- Distribution of community size is highly skewed
  - with power-law behavior in the upper tail
  - 90% have less than 65 members
    75% have less than 20 members



- Small communities (less than 10 members) are not popular.

Community size means… nothing?



The log-log plot of community size vs. number of bookmarks shows no correlation.

- Possible explanation: the amount of user activity is dictated by the nature of the community, not by the number of members.

RuSSIR
Russian Summer School in Information Retrieval

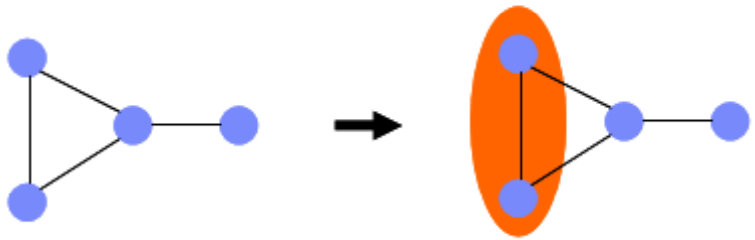
Clustering Algorithms

Cluster Analysis

- Hierarchical – agglomerative, divisive
- Iterative – exemplar-based (*k*-Means), commutation-based
- Density based – point concentration, cumulative attraction (MajorClust)
- Meta search controlled – gradient descent, competitive
- Statistical – Gaussian mixtures, …

Agglomeration

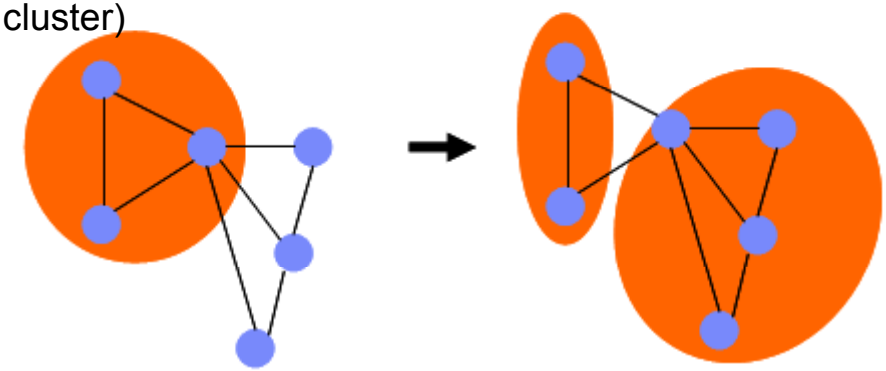
- Define majority decision (agglomeration)



Agglomeration

- Define majority decision (agglomeration)

- Define majority decision (node changes cluster)



K-means Algorithm (J. MacQueen, 1967)

The algorithm steps are:

- Choose the number of clusters, *k*.
- Randomly generate *k* clusters and determine the cluster centers (or directly generate *k* random points as cluster centers).
- Assign each point to the nearest cluster center, using a chosen distance measure.
- Recompute the new cluster centers.
- Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

Graph-theoretic distances and a median

Distances between two vectors
(cosine and Jaccard)

$$dist_{COS}(x, y) = 1 - \frac{x \bullet y}{\|x\| \cdot \|y\|} \ ,$$

$$dist_{JAC}(x, y) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} y_i^2 - \sum_{i=1}^{n} x_i y_i} \ .$$

where ***x*** and ***y*** are vectors representing two documents

K-means Algorithm (J. MacQueen, 1967)

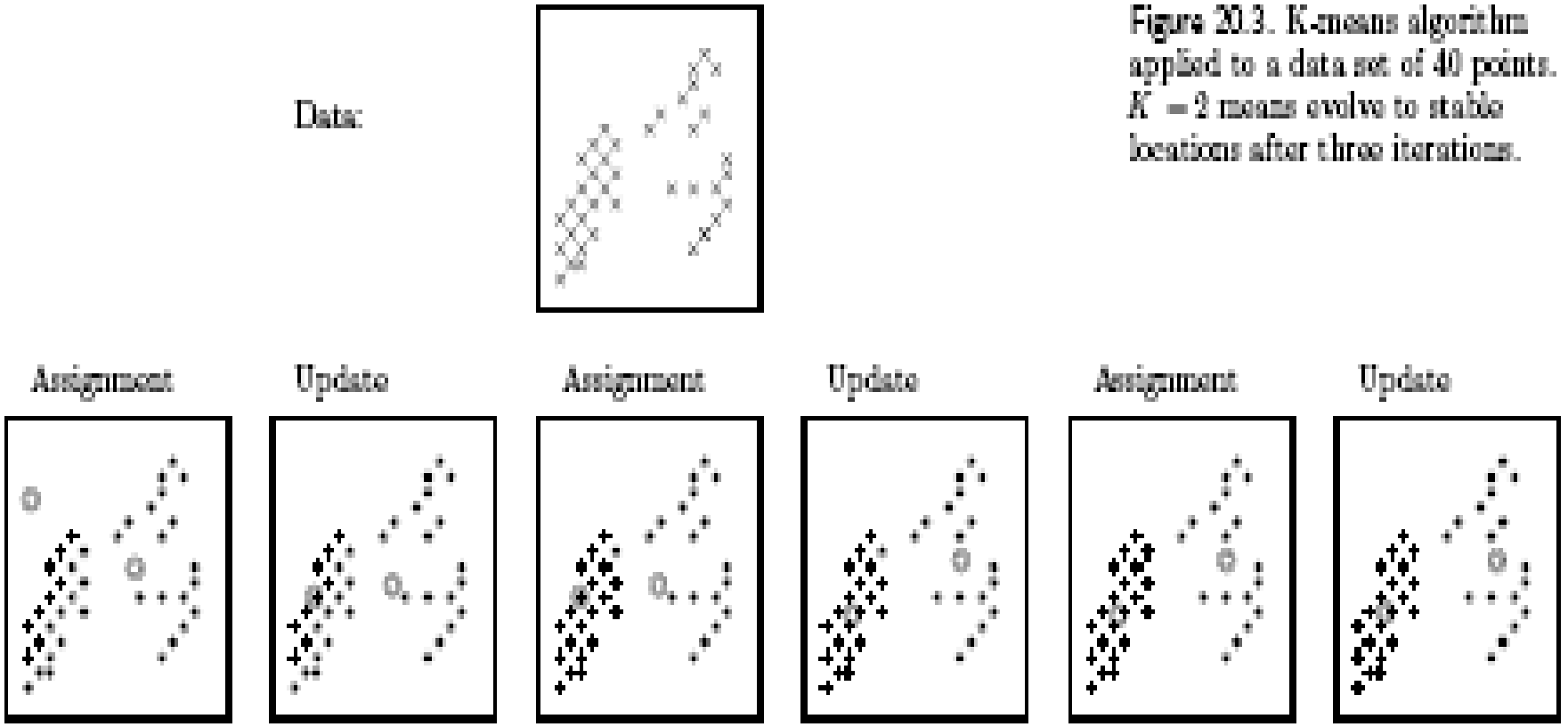


Figure 20.3. K-means algorithm applied to a data set of 40 points. $K = 2$ means evolve to stable locations after three iterations.

Picture is borrowed from the book by David McKay “Information Theory”. Cambridge University Press, 2003.

Graph-theoretic distances and a median

$$dist_{MCS}(G_1, G_2) = 1 - \frac{|mcs(G_1, G_2)|}{\max(|G_1|, |G_2|)},$$

where G_1 and G_2 are graphs, $mcs(G1, G2)$ is their maximum common subgraph, $\max(\cdots)$ is the standard numerical maximum operator, and $|\cdots|$ denotes the size of the graph. Usually, this is taken to be the sum of the number of nodes and edges in the graph, or the corresponding .node/edge frequency values.

Graph-theoretic distances and a median

- The ***median graph*** in a set of graphs is defined as a graph from the set which has the minimum average distance to all the other graphs in the set. Here the distance is computed as the graph-theoretic distance measure defined above.

K-means Graph Clustering Algorithm

(Schenker et al., 2005):

The algorithm steps are:

- Choose the number of clusters, *k*.
- Randomly generate *k* clusters and determine the graph centroids (= the median graphs) in each cluster or directly generate *k* random graphs as cluster centroids).
- Assign each graph to the **nearest** cluster centroid, using a chosen distance measure.
- Recompute the new cluster centroids.
- Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

Fuzzy c-Means Graph Algorithm

(Levner and Troussov, 2008, based on Bezdek ,1980)

Definitions and notation:

Fuzzy clustering is a process of assigning the membership levels to the elements (graphs), and then using them to assign data elements to one or more clusters.

For each graph *x* we have a coefficient $u_k(x)$ giving the degree of its being in the *k*th cluster. Usually, the sum of those coefficients over all clusters is defined to be 1.

In our **Fuzzy Graph *c*-Means Custering, the *fuzzy medan* (or fuzzy "*centroid*") of a cluster** is defined as a graph from the set which has the minimum **weighted** average distance to all the other graphs in the set. Here the **weighted distance** is computed as the graph-theoretic distance measure defined above multiplied by the coefficients $u_k(x)$.

Fuzzy c-Means Graph Algorithm (cont-d)

The degree of belonging x is the inverse of the *mcs*-distance between ***x*** and the median ***m*** of the cluster:

$$u_k(x) = 1/\, mcs(\boldsymbol{x}, \boldsymbol{m}) \qquad (1)$$

Then the coefficients are normalized so that their sum is 1.

Fuzzy c-Means Algorithm (cont-d)

Choose a number of clusters.

- Assign randomly to each graph coefficients $u_k(x)$ as the initial membership of being in the clusters.
- Compute the fuzzy median for each cluster, using the median definition and the formula (1) above.
- For each graph, compute its coefficients $u_k(x)$ of being in the clusters, using the formula (1) above.
- Repeat the two previous steps until the algorithm has converged (that is, the coefficients' $u_k(x)$ change between two iterations is no more than ε, the given *sensitivity threshold*).

The algorithm has the same problems as *k*-means, that is, its minimum is a local minimum, and the results depend on the initial choice of weights. **However, it successfully treats the butterfly effect.**

Λ-maximization problem (Stein&Niggemann, 1999)

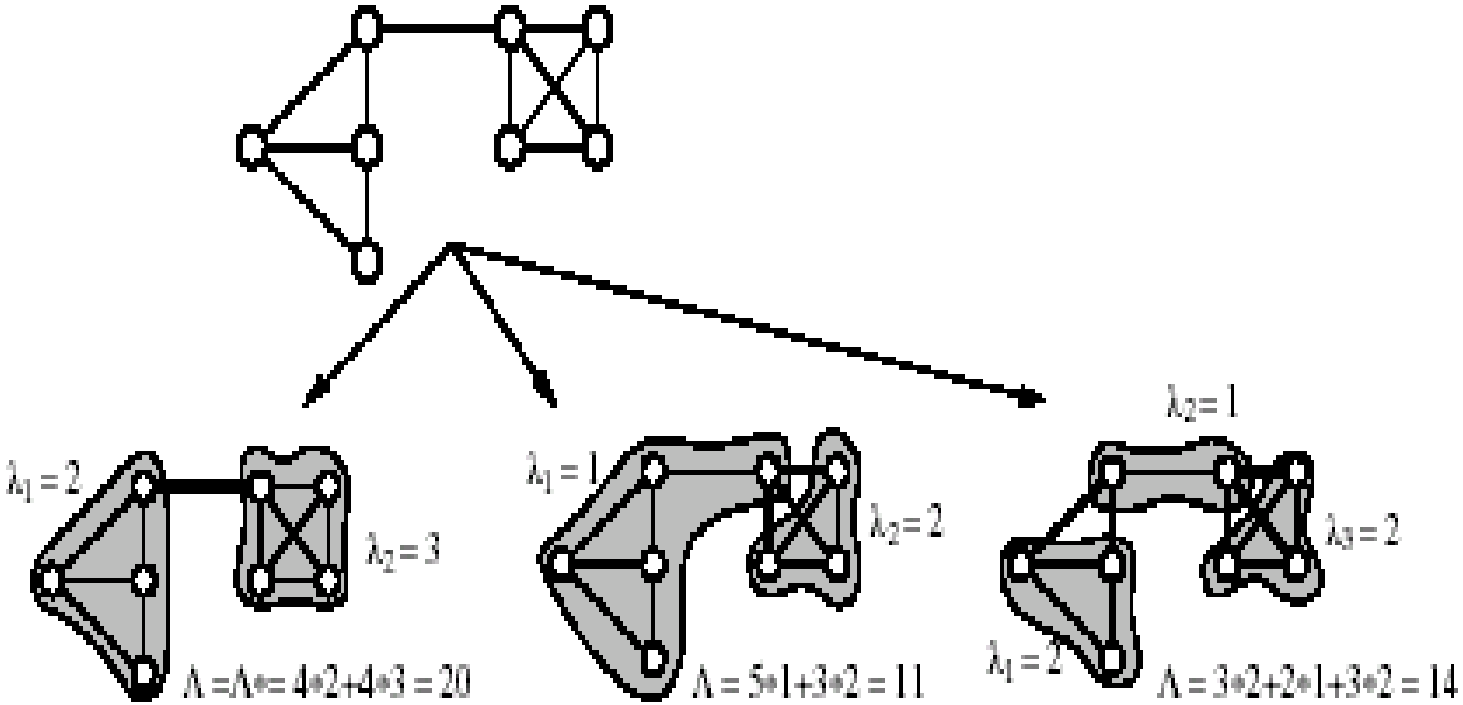
The problem is to rank the items (nodes) and find the cluster set maximizing the total cluster connectivity $\Lambda(C)$

The "cluster connectivity" is:

$$\Lambda(C) = \Sigma_{k=1,\ldots,K} |C_k| \lambda_k,$$

where C denotes the decomposition of the given graph G into clusters,
C_1, C_2, …, C_k are clusters in the decomposition C,
λ_k designates the edge connectivity of cluster $G(C_k)$, this is, the min number of edges that must be removed to make graph $G(C_k)$ disconnected,

MajorClust and Its Properties



Graph decomposition and related Λ(*C*) values
(adopted from [Stein&Niggemann, 1999])

MajorClust and the Stein-Niggemann Heuristic

-

The algorithm starts by assigning each “point” in the initial set its own cluster. Within the following re-labelling steps, a point adopts the same cluster label as the majority of its weighted neighbours.
If several such clusters exist, one of them is chosen randomly.
The algorithm terminates if no point changes its cluster membership

MajorClust and the Stein-Niggemann Heuristic

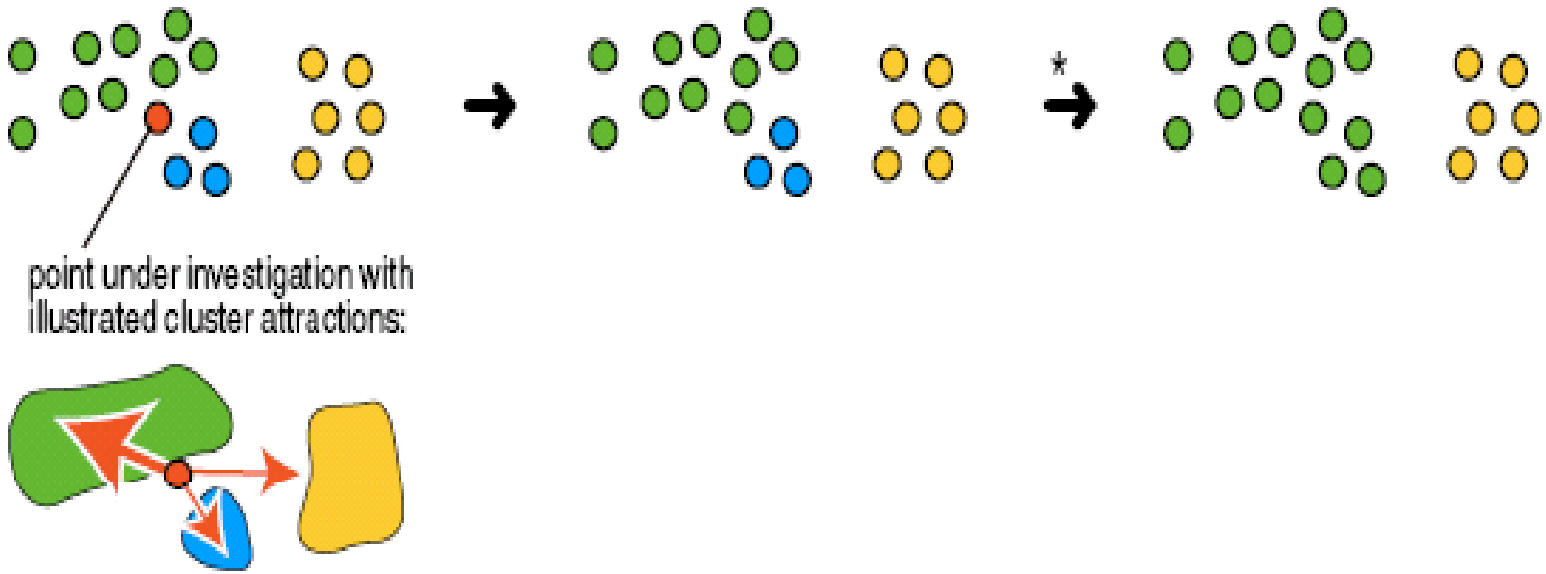


Figure 3. Illustration of MajorClust's clustering process; each cluster C exerts attraction to some point q depending on both its size, $|C|$, and distance to q.

(Figire borrowed from Stein&Busch, 2005)

The MajorClust Algorithm

Input: object set D, similarity measure φ,
: φ: D × D → [0; 1], similarity threshold t.
Output: function δ : D → N, which assigns a cluster label to each point.

Initialization (01) i := 0, *ready := false*

Iterations (02) for all p from D **do** i := i + 1, δ (p) := i **enddo**
(03) **while** *ready* = *false* **do**
(04) *ready := true*
Recomputation (05) for all q from D **do**
(06) δ* := i **if** Σ{φ(p, q)| φ(p; q) ≥ t and δ(p) = i} is maximum.
(07) **if** δ(q) ≠ δ* **then** δq) := δ*, *ready := false*
(08) **enddo**
(09) **enddo**

MajorClust and Its Properties

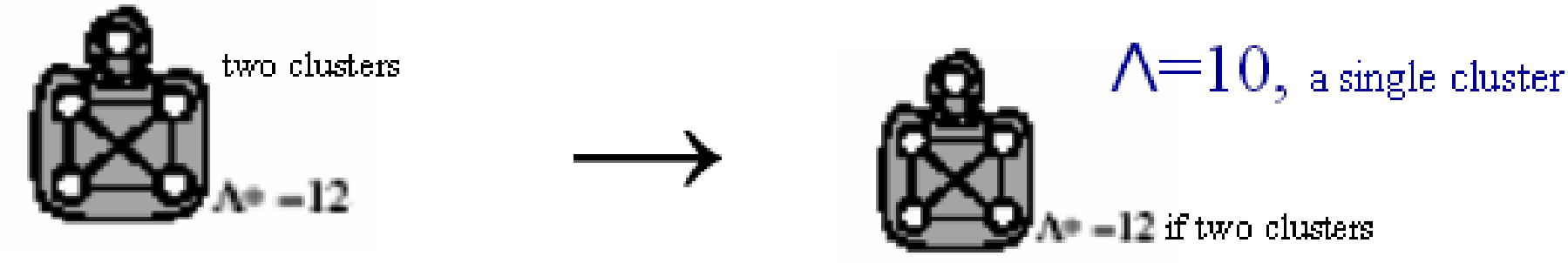
Advantages:

- It is simple,
- It is effective (fast and uses small memory), and
- It is locally optimal, providing a good ranking in practice.

MajorClust and Its Properties

Disadvantages

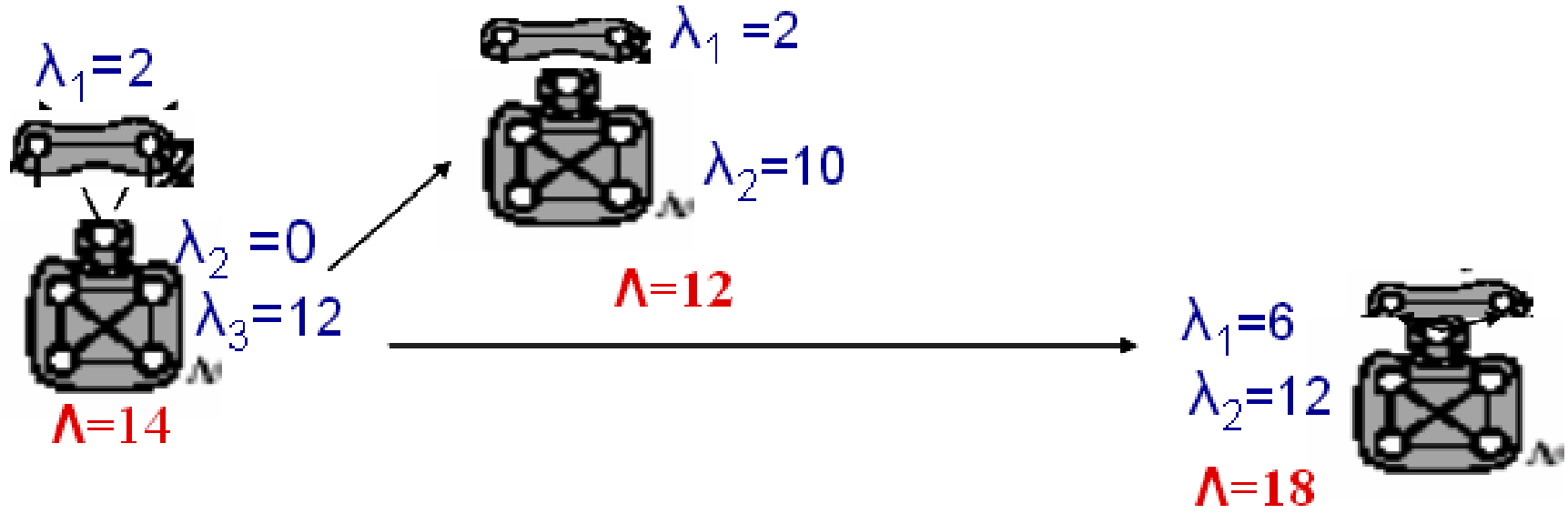
1. At a current step, MajorClust may insert into a cluster *a node with weak links*, i.e. the nodes with a small number of neighbours, which leads to the decrease of the objective function already achieved .



if two clusters

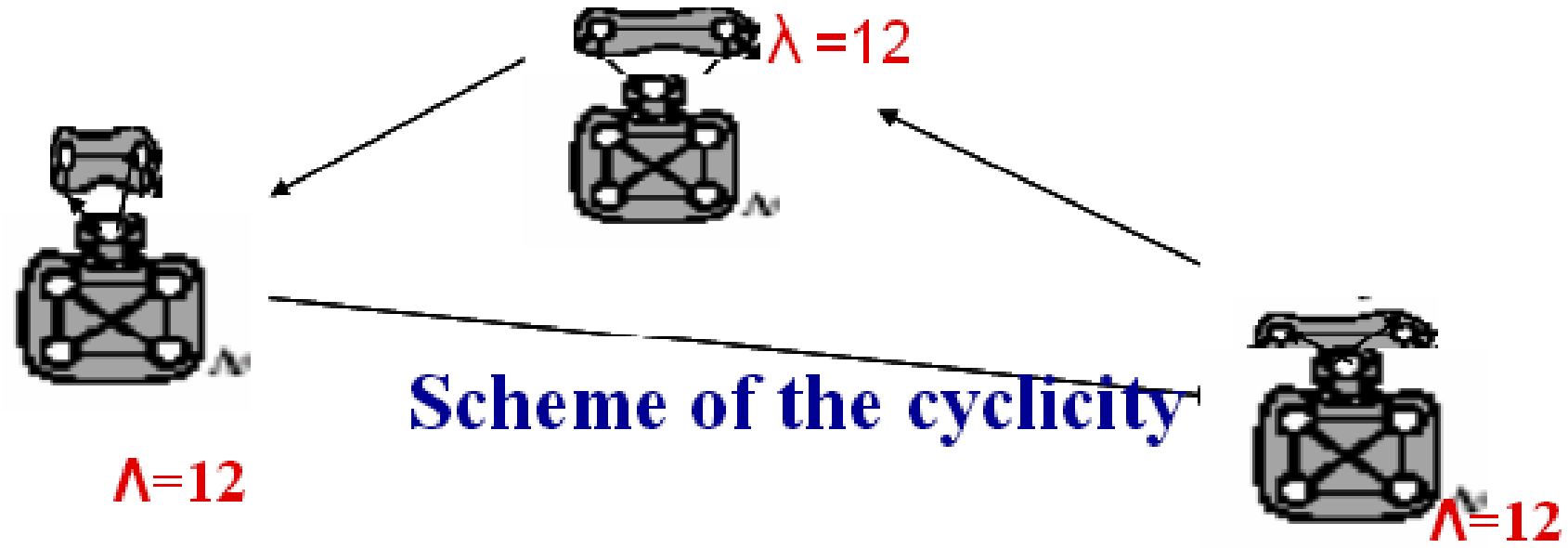
Disadvantages

2. If there are several “equivalent” candidate clusters with the same attraction for a current node, MajorClust assigns a node to one of the equivalent clusters in an arbitrary order, this may lead to the loss of a neighbouring good solution.



Disadvantages

3. In the case of ties, MajorClust scans the nodes of the graph in an arbitrary order, which may lead to the loss of good neighbouring solutions, as well as to the cyclicity.



Disadvantages

4. The simple MajorClust does not take into account different multiple attributes characterizing multi-dimensional links between the items (nodes) as well as the different contribution (weights) of different attributes.

MajorClust and Its Properties

Disadvantages

5. At each step, MajorClust takes into account only one local minimum ignoring many others which may lead to the loss of much better solutions than the one selected)
6. Major Clust assigns each target document to **exactly one cluster** though sometimes this it should belong to several clusters,

The above flaws will be avoided in the modified algorithm suggested, by the price of greater computational efforts (the running time) and a larger required memory.

FuzzyMajorClust and Its Properties

It what follows, the modification of the crisp MajorClust will be done in three directions:

- **1. Crisp modification of the basic MajorClust intended to avoid flaws caused by the ties and arbitrary choice of possible options in the basic MajorClust. This is done by introducing additional checks according to Rules R1-R5 (formulated below).**
- **2. Fuzzy modification FM1 of the basic MajorClust for a single attribute intended to treat the case when the items (documents) may belong to several categories. For this aim, the nodes and edges in graph are supplies by parameters called *membership functions*.**
- **3. Fuzzy modification FM2 of the MajorClust for the multiple attributes intended to treat the case when the items (documents) are characterized by contradicting attributes (and may belong to several different categories). This target is achieved by using the concept of *combined fitness* of items to categories and based on a weighted constrained SAS.**

FuzzyMajorClust and Its Properties

(Crisp) Modification of the Crisp MajorClust

- **Rule R1**. If there are several nodes having majority (or the maximum value of the corresponding objective function) in certain clusters, then *choose first the node* having the maximal number of the corresponding neighbours.

- **Rule R2**. If there are several nodes having both the majority and the maximum number of neighbours in certain clusters then *choose first* the node whose inclusion leads to the maximum increase of the objective function.

Modification of the crisp MajorClust (cont.)

- **Rule R3.** If, at some iterative step, the inclusion of some node would lead to the decrease of the objective function, this node should be skipped (that is, it will not be allocated into any new cluster at that step).
- **Rule R4**. If the inclusion of a single node is not attractive, try an inclusion of a group of two or three nodes.
- **Rule R5**. The algorithm stops when either :

1. the next iterative scanning of all nodes does not change the clustering (or does not improve its quality), or
2. the number of steps (or the running time) exceeds the given threshold.

FuzzyMajorClust and Its Properties

Fuzzy Modification FM1 of MajorClust

- The measure of membership of any edge *i* in a cluster *k* (=the weight) is presented by a membership function μ_{ik}, where $0 \le \mu_{ik} \le 1$, and $\Sigma_k \mu_{ik} = 1$ for any *i.*
- Definitions. A node *j* is called *interior* if all its neighbours belong to the same cluster as the node *j* itself. If an edge *i* connects nodes x and y, we will say that x and y are the *end nodes* of the edge *i*. A node *j* is called *boundary* if some of its neighbours belong to a cluster (or several clusters) other than the cluster containing the node *j* itself.

FuzzyMajorClust and Its Properties

Fuzzy Modification FM1 of MajorClust (cont.)

- We define $\mu_{ik} = a_{ik}/b_i$, where a_{ik} is the number of those neighbours of the end-nodes of edge *i* that belong to the same cluster *k* as the end-nodes of *i*, and b_i is the number of all neighbours to the end nodes of *i*.
- In a more advanced case of the *weighted graph*, we define

  $\mu_{ik} = A_{ik}/B_i$ where A_{ik} is the sum of the weights of edges linking the end-nodes of i with their neighbours that belong to the same cluster *k* as the end-nodes of i themselves*;* and B_i is the total sum of the weights of edges adjacent to the edge *i*.

FuzzyMajorClust and Its Properties

Fuzzy Modification FM1 of MajorClust (cont.)

- Further, we introduce the membership of **any item (node)** *j* in any cluster *k*, which is presented by another membership function, f_{jk} , where $0 \leq f_{jk} \leq 1$, and $\Sigma_k f_{jk} = 1$, for any j .

FuzzyMajorClust and Its Properties

Fuzzy Modification FM1 of MajorClust (cont.)

- In informal terms, the membership function of any edge i reflects the "*strength of connectivity*" between two nodes, the end-nodes of i, in a cluster, whereas the membership function of any node j reflects the semantic "*fitness*" (or "*attraction*") of node j with respect to a corresponding cluster *k* defined by the attributes of node j.
- For the both membership functions, it is valid:
  - $0 \leq \mu_{ik} \leq 1$, and $\Sigma_k \mu_{jk} = 1$, for any edge i , and

    $0 \leq f_{jk} \leq 1$, and $\Sigma_k f_{jk} = 1$, for any node j .

A New Fuzzy Λ-maximization Approach

$$\Lambda(C)=\Sigma_{k=1,\ldots,K}(\Sigma_{j=1,,|C^k|}\mu_{jk}\ \lambda_k+\Sigma_{j=1,,n}f_{jk}(C_k)),$$

where C denotes the decomposition of the given graph G into clusters,

C_1, C_2 , …, C_k are not-necessarily disjoint clusters in the decomposition C,

λ_k designates, as in MajorClust, the edge connectivity of cluster $G(C_k)$, that is, the min number of edges that must be removed to make graph $G(C_k)$ disconnected,

μ_{jk} is the (structural) membership degree of an edge containing node j in cluster k,

$f_{jk}(C_k)$ is the semantic fitness of node j to cluster k.

A New Fuzzy Λ-maximization Problem

The clustering problem is to find the cluster set maximizing the generalized cluster connectivity $\Lambda(C)$.

The *generalized cluster connectivity* is:

$$\Lambda(C)=\Sigma_{k=1,\ldots,K}(\Sigma_{j=1,\ldots,|C^k|}\mu_{jk}\lambda_k+\Sigma_{j=1,\ldots,n}f_{jk}(C_k)),$$

where C denotes the decomposition of the given graph G into clusters,

C_1, C_2, …, C_k are not-necessarily disjoint clusters in the decomposition C,

λ_k designates, as in the basic MajorClust, the edge connectivity of cluster $G(C_k)$, that is, the min number of edges that must be removed to make graph $G(C_k)$ disconnected,

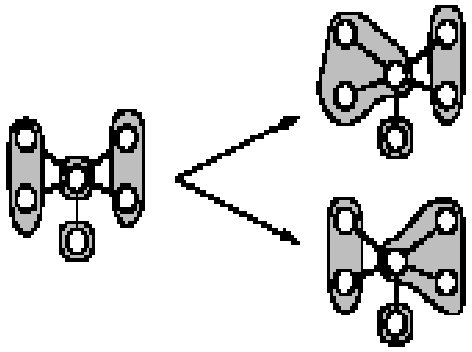
FuzzyMajorClust and Its Properties

The algorithm starts by assigning each point in the initial set its own cluster. Within the following re-labelling steps, a point adopts the same cluster label as the majority of its neighbours. If several such clusters exist, say z clusters, then a point adopts all of them and attains the membership function $\omega_{jk}=1/z$ for all clusters. At each step, if point j belongs to y clusters, then $\omega_{jk}=1/y$.

The algorithm terminates if no point changes its cluster membership

FuzzyMajorClust and Its Properties

- The fuzzy algorithm FuzzyMajorClust above preserves all advantages of the basic (crisp) MajorClust, avoids its flaws, and provides a natural belonging of an item to several clusters, when it follows from the structure (geometry) of the graph.



FuzzyMajorClust and Its Properties

- In what follows, we describe a multi-attribute version of the MajorClust formulated on a multigraph, each attribute being described by an edge of another color. The main idea is to integrate edge weights for multiple attributes into one integrated parameter called the *item-to-cluster (or page-to-cluster) fitness.*

Multi-Attribute FuzzyMajorClust and Its Properties

S1. Procedure for integrating the multiple attributes and finding f_{jk} consists of five stages:

- **First stage, identify: (a) *the set of attributes* characterizing the documents;**

  (b) *the current set of categories,* and

  (c) *the set of classifying words and expressions* within each category.

  These three sets can be either fixed, or flexible being extended or decreased during the interactive classification process.

Multi-Attribute FuzzyMajorClust and Its Properties

- **Second stage, compute the *fitness* measure f_{jkm} of document j to category k with respect to attribute *m*.**

The fitness f_{jkm} is a function of three arguments: (1) the number of words in the attribute m of document j that coincide with predefined classifying words and expressions in category k; (2) the size (the number of words) of attribute m of document j, and (3) the number of the classifying words in category k.

Multi-Attribute FuzzyMajorClust and Its Properties

$f_{jkm} = \min \{1, g_{jkm}/a_{km}\}$, where

g_{jkm} is # of words in m of doc j coinciding with training

(classifying) words of category k;

a_{km} the min required threshold of category k w. resp. to attribute m;

The integrated fitness of doc j to category k

$$f_{jk} = \sum_m w_m f_{jkm} ,$$

The total fitness of classification

$$F(C)=\sum_{j,k} v_k f_{jk} x_{jk} *(100/N)$$

Multi-Attribute FuzzyMajorClust and Its Properties

Example: Consider 5 attributes: *title, key words,abstract, bibliography, and authors' bio.*

$f_{jkm} = \min \{1, g_{jkm}/a_{km}\}$, where

g_{jkm} is # of words in attr. m of doc j coinciding with training

(classifying) words of category k;

a_{km} the min # of classifying words in cat. k with respect to attribute m whose presence in m provides the maximum fitness $f_{jkm} = 1$ (defined by experts);

Multi-Attribute FuzzyMajorClust and Its Properties

Third stage: The algorithm defines *weights* (relative importance) of categories v_j and *weights* w_k of attributes {*k*} which maybe either linguistic values (like, *very strong, strong, medium, weak, very weak*) or crisp magnitudes ranked in intervals [1, 100] or [0, 1].

Multi-Attribute FuzzyMajorClust and Its Properties

The fourth stage defines the complete fitness f_{jk} of document j to category k, by using an additive approach:

$f_{jk} = \sum_m w_m f_{jkm}$,

where the latter value can be either fuzzy or crisp.

Finally, **the fifth stage** distributes the documents among the categories using standard methods of cluster-analysis aimed either to maximize the total validity of classification or to maximize the total "fitness" of available documents to their assigned categories

$F(C) = (1/N) \sum_{j,k} v_k f_{jk} x_{jk} \times 100$,

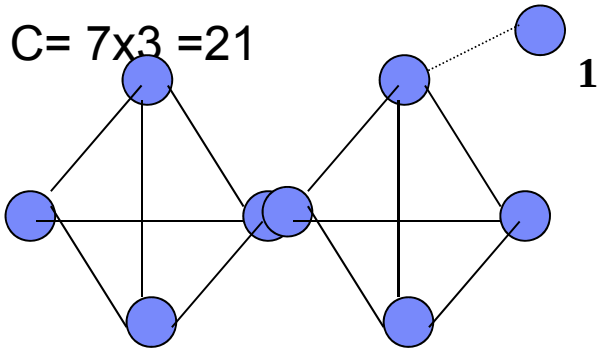
where $x_{jk} = 1$ if document j is assigned to category k and 0 otherwise,

and N is the number of documents, under predetermined constraints on the cardinality of category sets and running time of the classifying procedure.

Example

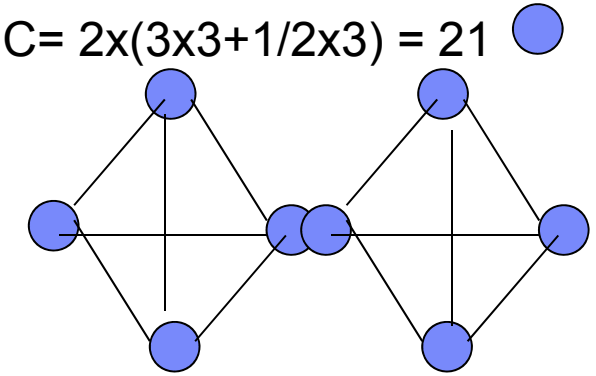
Two clusters (counting 1);

C= 7x3 =21



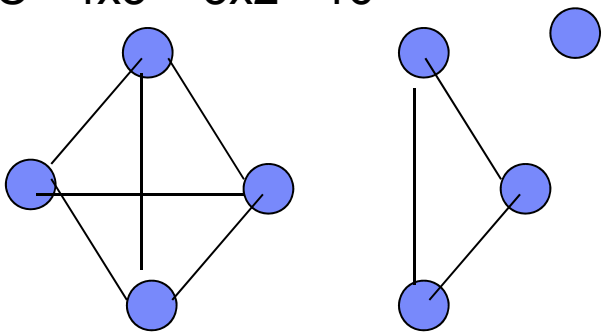
Three clusters (two overlapping);

C= 2x(3x3+1/2x3) = 21

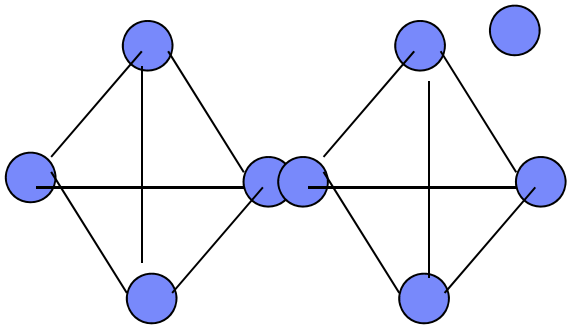


Three clusters (crisp);

C= 4x3 + 3x2 =16



Three clusters (two overlapping), with semantic grading; C= 2x(3x3+1x3) = 24



RuSSIR
Russian Summer School in Information Retrieval

Spreading Activation Methods

Spreading Activation (as search)

- The search process is initiated by labeling a set of source nodes (e.g. concepts in a semantic network) with weights or "activation" and then iteratively propagating or "spreading" that activation out to other nodes linked to the source nodes. Most often these "weights" are real values that decay as activation propagates through the network. Activation terminates when two alternate paths reach the same node.

SAS – Informal description

Initialize by setting all activation values A [i] = 0. Set one or more origin nodes to an initial activation value A [i] greater than the firing threshold F. A typical initial value is 1.0.

- For each unfired node [i] having an a.v. A [i] > the threshold F:
- For each Link [i, j] connecting the source node [i] with target node [j], adjust

  A [j] = A [j] + (A [i] * W [i, j] * D) where D is the decay factor.
- If a target node receives its a.v. >1.0, then set it to 1.0. Maintain a.v. = 0.0 as a lower bound on the target node's activation value.
- Nodes receiving a new activation value that exceeds the firing threshold F are marked for firing on the next s.a. cycle.
- The procedure terminates when either there are no more nodes to fire or in the case of marker passing from multiple origins, when a node is reached from from more than one path.
- Variations of the algorithm that permit repeated node firings and activation loops in the graph, terminate after a steady activation state, with respect to some delta, is reached, or when a maximum number of iterations is exceeded, or a target state is reached, or there are no more nodes in the priority queue

Formal description of SAS

A network is modeled by a directed graph $G = (V,E)$ where

- V is the set of vertices v_i, and E is the set of edges e_j
- $imp(v)$, $Imp(e)$ is the importance value of arcs and nodes.
- w – "weights" of links, $0 \leq w \leq 1$.
- $A(V)$ – is the "activation" function. Usually, a real valued function on nodes of the network.

The SAS Algorithm

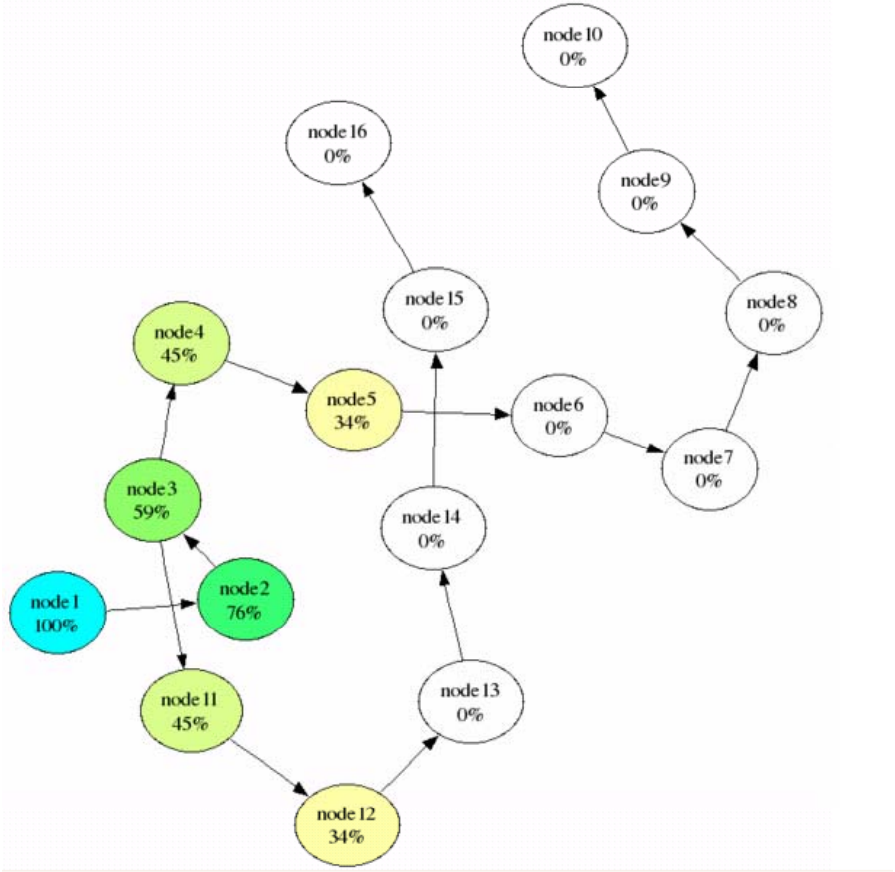
- Initialisation
  Sets A [v_i] = 0 for all nonactivated nodes in V.
- Sets A [v_i] = F (usually =1) for one or more origin nodes in V.
- Iterations
  - Recomputation

    Input (*e*) = F(v) · (1/outdegree(v)**beta) (where v=tail(*e*))
    Output (*e*) = Input(*e*) · D(e) (where D is decay)
  - Activation of node v through incoming links *e*::
  - Input(*v*) = Σ_e Output(*e*) for all incoming links e
  - New level of activation $F_{new}(v) = F(v) + Input(v)$
  - List Purging
    Exclude the nodes with the values less than a threshold.

  - Checking Constraints
    like maximum number of iterations to be performed.
- Output
  The list of nodes (with values after SAS) ranked according to *F* values.

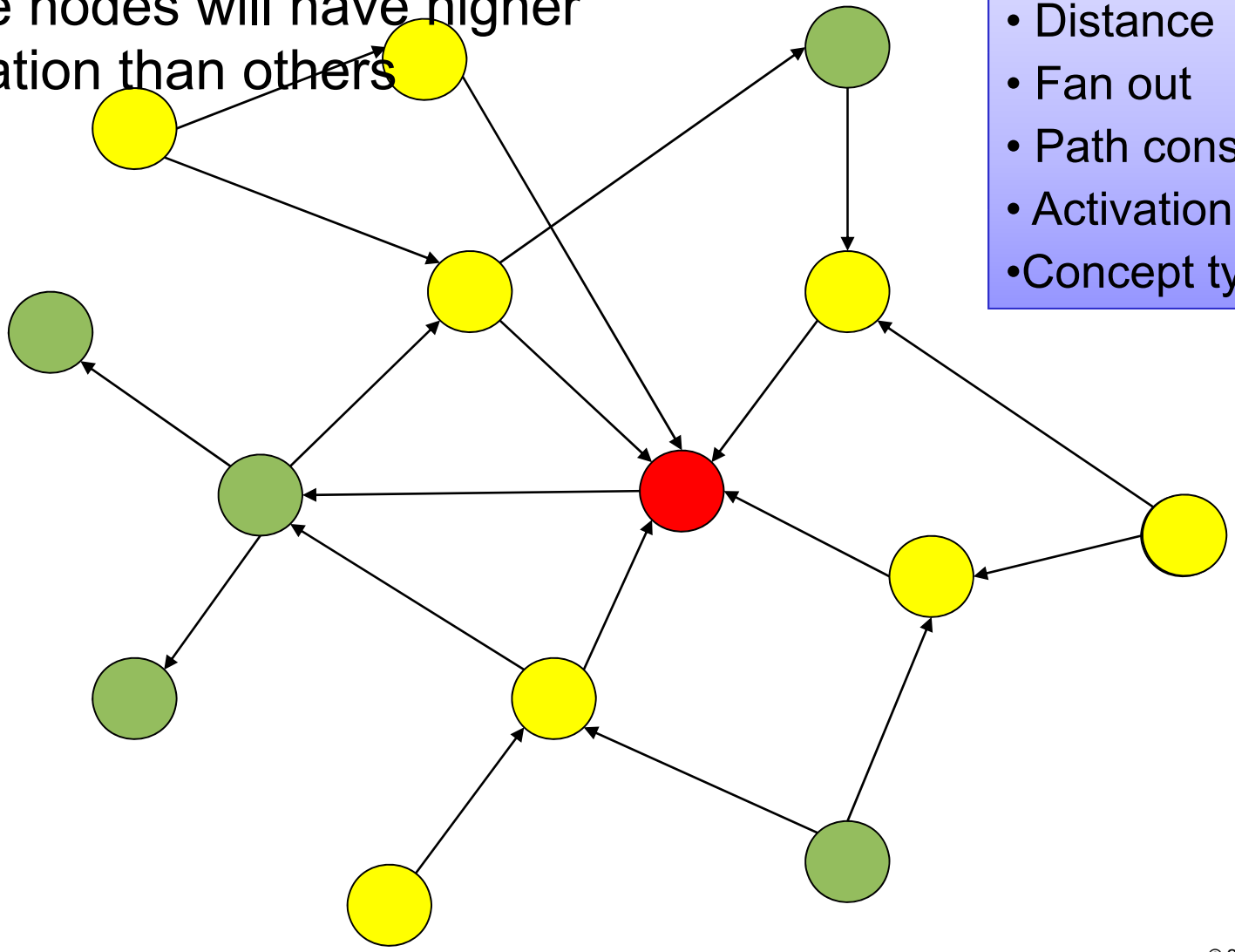
Spreading activation originated at node 1 (blue) which has an initial activation value of 1.0 (100%). Each link has the same weight value of 0.5. The decay factor was 0.85. Four cycles of spreading activation have occurred. Color hue and saturation indicate different activation values.

(borrowed from wikipedia http://en.wikipedia.org/wiki/Spreading_activation)



The Constrained SAS

Some nodes will have higher activation than others



Constraints

- Distance
- Fan out
- Path constraints
- Activation threshold
- Concept type

SAS with Weights

- Subsumption Weight
- Path Length Weight
- Context Weight
- Grammatical structure weight
- Semantic structure weight
- Trust Weight