Node Similarities from Spreading Activation

Kilian Thiel and Michael R. Berthold

Nycomed Chair for Bioinformatics and Information Mining
Dept. of Computer and Information Science
University of Konstanz, Konstanz, Germany
Kilian.Thiel@Uni-Konstanz.DE

Abstract. In this paper we propose two methods to derive different kinds of node neighborhood based similarities in a network. The first similarity measure focuses on the overlap of direct and indirect neighbors. The second similarity compares nodes based on the structure of their possibly also very distant neighborhoods. Both similarities are derived from spreading activation patterns over time. Whereas in the first method the activation patterns are directly compared, in the second method the relative change of activation over time is compared. We applied both methods to a real world graph dataset and discuss some of the results in more detail.

1 Introduction

It is essential for many experts of various fields to consider all or at least the bigger part of their accessible data before making decisions, to make sure that no important pieces of information are ignored or underestimated. In many areas the amount of available data grows rapidly and manual exploration is therefore not feasible. Many of these datasets consist of units of information, e.g. genes, or proteins in biomedical datasets, or terms and documents in text datasets, as well as relations between these units, and thus can be represented as networks, with units of information represented as nodes or vertices and their relations as edges.

For analysts and experts it can be interesting to find nodes in such networks that are directly or indirectly connected to given query nodes in order to find a community, or a dense subgraph located around a given query node. In terms of biomedical networks, proteins can be found interacting with a query protein, or sharing certain properties. In social networks a circle of friends or acquaintances of a person can be determined, and in textual networks frequently shared terms or documents, according to a query can be discovered. To identify and extract closely connected nodes to certain query nodes, often methods based on spreading activation are used, especially in the field of information retrieval [9,10,4].

Besides finding nodes, which are part of the community of a query node and thereby closely positioned to it, the discovery of structurally similar nodes can be desirable, too. Nodes are structurally similar if the connection structure of their neighborhoods is similar. An overlap of neighborhoods is not required, which means that the nodes can be located far away from each other. For instance structurally similar individuals in social networks may play the same role in their community. In biomedical networks, for

M.R. Berthold (Ed.): Bisociative Knowledge Discovery, LNAI 7250, pp. 246–262, 2012.

example, proteins can be found playing the same role in their metabolic pathways. Additionally the comparison of communities of the query node and its structurally similar result nodes, can lead to new insights as well. For instance information such as the number and size of sub-communities, the number and connectedness of central nodes, and the structural position of the query, and result nodes in their corresponding community, can be interesting.

Experts and analysts do not always know exactly what to look for, or where. Thus the usage of classical information retrieval systems, requiring specific queries, is often not sufficient. Methods that suggest unknown, interesting and potentially relevant pieces of information around a certain topic can help to find a focus, induce new ideas, or support creative thinking. In [11,18] these pieces of information are described as domain bridging associations, or *bisociations*. The underlying data is thereby organized in a *bisociative network* or *BisoNet*, consisting of units of information and their relations [6,17]. A bisociation pattern based on the structural similarity of two subgraphs from different knowledge domains is defined in [17,18].

In contrast to nodes of bisociations based on bridging concepts and bridging graphs, nodes of bisociations based on structural similarity do not necessarily have to be positioned close to each other. These patterns of bisociation link domains, which may not have any direct connections by means of the abstract concepts they have in common, are represented by the structure of their node neighborhood. A prodrug that passes the blood-brain barrier by carrier-mediated transport, and soldiers who pass the gate of Troy hidden in a wooden horse are examples of the kind of bisociation, described in [18]. Both the prodrug as well as the soldiers cannot pass the barrier or gate without the help of a carrier. The abstract concept of using a carrier in order to pass a barrier is represented by the structure of the corresponding subgraph. Finding nodes that are structurally similar to a query node, extracting densely connected, direct and indirect neighbor nodes, and comparing these subgraphs can lead to the discovery of structural bisociations.

Therefore two different kinds of node similarities can be used, structural and spatial similarity. We propose two methods to derive these two kinds of similarities between nodes in a graph from spreading activation processes. The first method is based on the comparison of activation vectors, yielding a spatial similarity. The second method is based on the comparison of change of activation, the *velocity*, yielding a structural similarity. The focus of this article is the definition and explanation of these two similarities and their application to a real world dataset in order to estimate their suitability.

The article is organized as follows. The next section concerns related work about spreading activation and node similarities. Section 3 defines the preliminaries of spreading activation processes on graphs and the underlying framework. The concept of signature vectors, which can be derived from spreading activation processes to represent nodes is introduced in Section 4. In Section 5 we introduce two kinds of node similarities based on the comparison of activation vectors and signature vectors. This is followed by Section 6, which describes the application of these similarities on the Schools-Wikipedia[1] (2008/09) data. Finally Section 7 concludes the article.

[1] `http://schools-wikipedia.org/`

2 Related Work

In the field of graph analysis different kinds of node equivalences, especially in the context of role assignments have been made [8]. Thereby nodes can be considered equivalent based on different properties, such as neighborhood identity, neighborhood equivalence, automorphic mapping, equal role assignments to neighboring nodes, and others. Networks from real world data are usually noisy and irregular, which makes finding equivalent nodes unlikely. Thus a relaxed notion of equivalence, in the sense that nodes are defined similarly to a certain extent, based on certain properties, is useful for a robust comparison of nodes [20].

Approaches which are conceptually similar to the comparison of activation pattern of nodes, from spreading activation processes are given in [22,23,19]. These approaches, like spreading activation, base on an iterative process, consider nodes to be more similar the more their direct and indirect neighborhood overlaps. The aim of these approaches is to detect dense clusters and communities. Since they also take into account an overlap of indirect node neighborhoods as well, they are more robust than measures comparing only the direct neighborhoods, such as e.g. the *Jaccard index* [14].

Each of these approaches suffers from different drawbacks. In [22] the characteristic node vectors of the corresponding normalized adjacency matrix are projected onto a lower dimensional space. Then the values of the projected vectors of each node are replaced iteratively by the mean of their neighbor values. Due to the projection into a lower dimensional space, information can get missing. In [23] node distances are determined, based on random walks, which are iterative processes as well. Nodes are similar if the probability of reaching other nodes in a specified number of iterations is similar. Here only walks of a certain length are considered when computing the distances. However, a more general variant of the algorithm considers walks of different lengths as well. Taking into account all computed iterations, as in [19] may yield to higher accuracy. In [19] all iteration results are accumulated with a decay to decrease the impact of the global node neighborhood. Since the accumulated and normalized activation values are used as similarities the method may yield asymmetric similarities on directed graphs.

In our approach we compare the computed activation pattern by means of a well known similarity measure, the *cosine similarity*, yielding symmetric values. Thus our method can be applied to directed graphs as well. We do not use a lower dimensional node representation by means of a projection into a lower dimensional space and hence may not lose information. We consider all iteration results up to a maximal number of iterations and not only walks of a certain length.

Additionally we propose a second node similarity derived from the comparison of activation changes in each iteration. Based on this method nodes are similar if the structure of their neighborhood is similar, although the neighborhood does not need to overlap at all. This yields a completely different similarity compared to those mentioned above.

Originally spreading activation was proposed by Quillian [24,25] and Collins et al. [9] to query information networks. The method facilitates the extraction of subgraphs, nodes and edges directly and indirectly related to a given query. Initially the nodes representing the query are activated. The activation is than spread iteratively to

adjacent nodes, which are activated with a certain level as well until a termination criterion is reached or the process converges. The subset of activated nodes, their level of activation, as well as the induced subgraph compose the final result. The level of activation of nodes is often used as relevancy heuristic.

Spreading activation has been applied in many fields of research from semantic networks [25], associative retrieval [27], to psychology [9,1,2], information retrieval [26,4,10,3] and others [13,28,21,16,12]. Most of these approaches use a set of common heuristic constraints [10] in order to restrict the dynamics of the process, such as distance constraints to terminate the process after a certain number of iterations, or fan out constraints to avoid excessive spreading. In [5] it is shown that pure (constraint free) spreading activation with a linear activation function on a connected and not bipartite graph always converges to the principal eigenvector of the adjacency matrix of the graph.

Usually the level of activation itself, which is sometimes normalized or accumulated over the iterations, represents the relevancy or similarity of nodes to a given query. We propose the comparison of (accumulated) activation patterns, as well as the change of activation patterns to determine similarities between nodes of the underlying network.

In the next section the preliminaries of spreading activation and its framework, that we use in this work are defined.

3 Spreading Activation

Activation is spread on a graph $G = (V, E, w)$, with V as the set of nodes $V = \{1, \ldots, n\}$, $E \subseteq V \times V$ as the set of edges and $w(u, v)$ as the weight of the edge connecting u and v, with $u, v \in V$, $w(u, v) = 0$ if $(u, v) \notin E$. For an ease of exposition we assume that the graph G is undirected, however our results easily generalize to directed graphs. The activation state at a certain time k is denoted by $\mathbf{a}^{(k)} \in \mathbb{R}^n$ with $\mathbf{a}_v^{(k)}$ as the activation of node $v \in V$. Each state $\mathbf{a}^{(k)}$ with $k > 0$ is obtained from the previous state $\mathbf{a}^{(k-1)}$ by the three families of functions described below.

- *Input function*: combines the incoming activation from adjacent nodes.
- *Activation function*: determines the state of activation based on the incoming activation.
- *Output function*: determines the outgoing activation based on the current activation.

The initial state $\mathbf{a}^{(0)}$ defines the activation of nodes representing the query. In each iteration activation is spread to adjacent nodes activating them with a certain level as well. The process is usually terminated after a certain number of iterations, activated nodes or convergence.

3.1 Linear Standard Scenario

In our approach we use a linear standard scenario described in [5] for which convergence is shown for non-bipartite connected graphs. The input, activation, and output

function can be combined to one function. Given a graph $G = (V, E, w)$ and an activation state $\mathbf{a}^{(k-1)}$ at time $k-1$, the activation of a certain node v at time k is defined by

$$\mathbf{a}_v^{(k)} = \sum_{u \in N(v)} w(u,v) \cdot \mathbf{a}_u^{(k-1)}, \forall v \in V, \tag{1}$$

with $N(v) = \{u \,|\, \{u,v\} \in E\}$ as the set of neighbors of v. Furthermore the spreading activation process can be described in matrix notation. With $W \in \mathbb{R}^{n \times n}$ as the weight matrix defined by $(W)_{uv} = w(u,v)$ a single iteration can be stated as $\mathbf{a}^{(k)} = W\mathbf{a}^{(k-1)}$ leading to

$$\mathbf{a}^{(k)} = W^k \mathbf{a}^{(0)} \,. \tag{2}$$

Note that this holds for undirected graphs only. In general an iteration can be stated as $\mathbf{a}^{(k)} = (W^T)^k \mathbf{a}^{(0)}$, holding for directed graphs as well. In order to prevent the activation values from increasing heavily or vanishing, the activation vector is normalized by its Euclidean length after each iteration.

$$\mathbf{a}^{(k)} = \frac{W^k \mathbf{a}^{(0)}}{\left\| W^k \mathbf{a}^{(0)} \right\|} \,. \tag{3}$$

Rescaling does not change the direction of the activation vector, so convergence to the principal eigenvector $\mathbf{v}_1$ of W is still ensured since $\lim_{k \to \infty} \mathbf{a}^{(k)} = \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|}$.

4 Node Signatures

Convergence of the spreading activation process yields to query independent results. No matter from which node(s) spreading processes have been started initially, the activation state becomes equal after a sufficient number of iterations. From iteration to iteration, activation vectors change their directions towards the direction of the principal eigenvector of the weight matrix W. How quickly a process converges can be described by its velocity and depends on the node(s) from which it was started. For each node the corresponding convergence speed can be determined and represented as a vector, called *signature vector*.

The velocity represents the change of direction of activation patterns between each subsequent iterations. A velocity vector at time k of a spreading process started at $v \in V$ is defined as

$$\delta^{(k)}(v) = \begin{cases} \mathbf{0} & \text{, if } k = 0 \\ \mathbf{a}^{(k)}(v) - \mathbf{a}^{(k-1)}(v) & \text{, else} \end{cases} , \tag{4}$$

with $\mathbf{0}$ as a vector of all 0 and $\mathbf{a}^{(k)}(v)$ as the activation vector at iteration k of a spreading process started at node v, whereas

$$\mathbf{a}_i^{(0)}(v) = \begin{cases} 1 & \text{, if } i = v \\ 0 & \text{, else} \end{cases} ,$$

for all $i \in V$. A norm of a velocity vector represents the amount of change, the step size of the process towards the principal eigenvector of the adjacency matrix. In this work

we use the l_2 norm as step size $\|\cdot\|$. Based on the step sizes of each iteration k up to a maximum number of iterations k_{max}, with $0 \leq k \leq k_{max}$, the signature vector of each node is defined. This vector provides information about the convergence speed of a spreading process, starting from a certain node v and is defined as

$$\tau_k(v) = \left\| \delta^{(k)}(v) \right\|, \tag{5}$$

with $\tau(v) \in \mathbb{R}^{k_{max}}$.

5 Node Similarities

Two kinds of node similarities can be derived based on the comparison of activation and convergence behaviors of spreading activation processes starting from each node. On the one hand nodes can be considered similar if their activation vectors are similar (*activation similarity*). On the other hand nodes can be considered similar if the change of activation from one iteration to another is similar (*signature similarity*).

These two kinds of similarities compare nodes based on two different properties, (direct and indirect) neighborhood overlap or neighborhood similarity. A neighborhood overlap between two nodes means that a part of the neighborhood of these two nodes is identical. This consequently means, the larger the overlap the closer the nodes are in the graph. This property yields a spatial similarity measure and is taken into account when activation vectors are compared (activation similarity). A neighborhood similarity of two nodes means that their neighborhood is structurally equivalent to a certain degree but not necessarily identical [20], which can be determined when comparing the change of activation vectors (signature similarity). This property yields a structural similarity measure.

Two node partitionings based on these two different properties are illustrated in Figure 1. The partitioning is indicated by the shading of the nodes. Nodes with the same shade are considered maximally similar (with a similarity value of 1) w.r.t. an equivalent (Figure 1a) or identical (Figure 1b) neighborhood. In Figure 1a the white as well as the black nodes are structurally equivalent since they are automorphic images of each other [8]. In Figure 1b the leaf nodes $\{4, 5, 6, 7\}$, $\{8, 9, 10, 11\}$ and $\{12, 13, 14, 15\}$ are the most similar nodes, due to their identical neighborhood, depicted by the shading gray, black, and white. Even if the leaf nodes are structurally equivalent only those with an identical neighborhood are highly similar. Furthermore the three nodes in the middle $\{1, 2, 3\}$ are not equal based on the comparison of their neighborhood. Node 3 is more similar to $\{12, 13, 14, 15\}$ than to 1 or 2 when comparing their pattern of activation.

The two different similarity measures derived from spreading activation processes allow on the one hand for the identification of structurally similar nodes to a given query node, even if they are located far apart in the graph via the signature similarity. On the other hand a densely connected subgraph of direct and indirect neighbors can be extracted for each node applying the activation similarity measure. In the following, these two node similarities are formalized and described in detail.

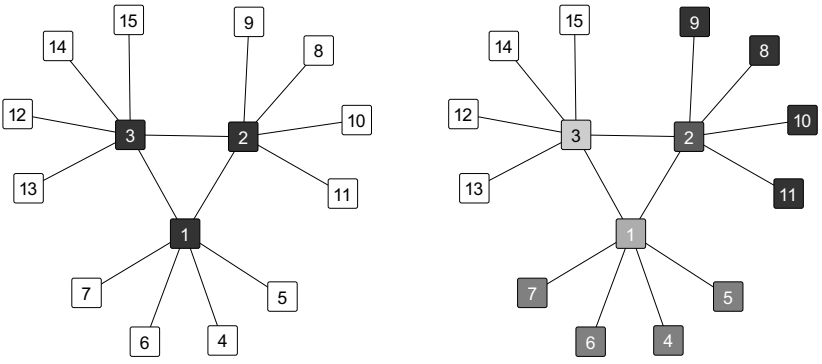


Fig. 1. Two node partitionings, indicated by the shading based on two different node properties, equivalent and identical neighborhood. In 1a the white nodes are structurally equivalent as well as the black nodes, which can be determined by the comparison of the signature vectors (signature similarity). In 1b the leaf nodes are divided into three partitions white, gray, and black, since their neighborhood is only partially identical. In addition node 3 is more similar to the white nodes, node 2 to the black nodes, and node 1 to the gray nodes than to others, which can be determined by the comparison of the accumulated activation vectors (activation similarity).

5.1 Activation Similarity

The first similarity described is based on the comparison of activation vectors and named *activation similarity*. The sequence of activation states of a spreading process started from a certain node describes the node relative to its local and global neighborhood in the graph. Dependent on its neighborhood many or few nodes will be activated and activation will spread fast or slow. Nodes close to the initially activated node will get activated sooner than nodes further apart from this node. Furthermore nodes will get activated to a higher level, at least in the primary iterations, if many walks of different lengths exist, connecting these nodes with the initially activated node. Nodes that are similarly connected to a shared neighborhood will induce similar activation states.

The level of activation $\mathbf{a}_i^{(k)}(v)$ of a node $i \in V$ at a time k, induced by a spreading process started at node v, reflects the reachability of i from node v along (weighted) connecting walks of length k. The more (highly weighted) walks of length k exist connecting i and v, the higher the level of activation. A query node u inducing a similar level of activation $\mathbf{a}_i^{(k)}(u)$ at node i at iteration k is consequently similarly connected to i along (weighted) connecting walks of length k.

Comparing the activation pattern of iterations $k \geq 1$ allows for the determination of the direct and indirect neighborhood overlap of nodes, whereas measures like the *cosine similarity*

$$\sigma_{\cos}(u, v) = \frac{|N(u) \cap N(v)|}{\sqrt{|N(u)|\,|N(v)|}}$$

or *Jaccard index* [14]

$$\sigma_{\text{jaccard}}(u, v) = \frac{|N(u) \cap N(v)|}{|N(u) \cup N(v)|}$$

based on the characteristic vectors of nodes allow for a comparison of the direct node neighborhood only. Figure 2 depicts a graph for which the cosine node similarities have been computed and indicated by the node shading. The nodes 1 and 3 have a cosine similarity of 1, since their direct neighborhood (node 2) is identical. The nodes 1 and 2 have a similarity of 0, as well as nodes 2 and 3. Although they are direct neighbors, the related similarity is 0 since the particular direct neighborhoods do not overlap. Consideration of the indirect neighborhood via connecting walks of lengths greater than 1 ($k > 1$) by applying activation similarity (with 5 iterations) still yields a similarity of 1 for nodes 1 and 3 due to their identical neighborhood, but a similarity greater than 0 for nodes 1 and 2 as well as for nodes 2 and 3. For the detection of dense subgraphs, comparison of the direct node neighborhoods only is too strict. Not all of the nodes in a dense subregion necessarily share a direct neighborhood. Taking into account the indirect k-neighborhoods yields a more robust similarity measure.

Fig. 2. Nodes 1 and 3 (white) have a cosine node similarity of 1, since their direct neighborhood is identical. Although nodes 1 and 2, as well as 2 and 3 are direct neighbors their cosine similarity is 0 since their direct neighborhood is not overlapping.

In [23] it is stated that in terms of random walks of length k starting from a node v the probability is high for other nodes to be reached if they are located in the same dense subgraph or community as node v. For an additional node u, the probability of reaching these nodes is high as well if it is located in the same community. Since random walks are driven by power iterations of the transition matrix of a graph they can be seen as spreading activation processes on a normalized weight matrix.

Considering not only walks of a certain length k as in [23] but all connecting walks of different lengths as in [19] provides a more detailed representation of the local and global neighborhood of a node. Accumulating all activation vectors $\mathbf{a}^{(k)}(v)$ from a spreading process starting from v with a decay α results in a final activation vector $\mathbf{a}^*(v)$ defined by

$$\mathbf{a}^*(v) = \sum_{k=1}^{k_{max}} \alpha^k \mathbf{a}^{(k)}(v) \tag{6}$$

with $0 < \alpha < 1$. The decay α decreases the impact of longer walks and ensures convergence for $k_{max} \to \infty$ for l_2 normalized systems [5]. It is reasonable to decrease the contribution of longer walks in order to keep more information about the local neighborhood of v. The above mentioned form is closely related to the centrality index of Katz [15]. We do not want to let the series fully converge since activation vectors of latter iterations do not contribute much to the final activation based on the decay α, and become more and more similar due to convergence of the spreading processes. We

chose k_{max} based on the convergence behavior of the underlying graph as well as the decay factor α.

Before a similarity on the final activation vectors is defined it needs to be considered that nodes with very high degrees will be activated to a higher level. They are more likely to be reached even if they are not located in the same dense region as the node from which activation has spread initially. To take this into account we normalize, related to [23], the final activation by the degree of the corresponding node. The degree normalized final activation vector is thereby denoted as

$$\hat{\mathbf{a}}^*(v) = D^{-\frac{1}{2}}\mathbf{a}^*(v) = D^{-\frac{1}{2}}\left(\sum_{k=1}^{k_{max}} \alpha^k \mathbf{a}^{(k)}(v)\right) \tag{7}$$

with D as the (weighted) degree matrix defined by $(D)_{ii} = d(i), (D)_{ij} = 0$ for $i \neq j, \forall i$ and $d(i) = \sum_{j=1}^{n} (W)_{ij}$. Based on these normalized final activation vectors we define the activation similarity between two nodes u and v

$$\begin{aligned} \sigma_{\text{act}}(u,v) &= \cos(\hat{\mathbf{a}}^*(v), \hat{\mathbf{a}}^*(u)) \\ &= \frac{\langle \hat{\mathbf{a}}^*(u), \hat{\mathbf{a}}^*(v)\rangle}{\|\hat{\mathbf{a}}^*(u)\|\,\|\hat{\mathbf{a}}^*(v)\|} \\ &= \frac{\sum_{i=1}^{n} \mathbf{a}_i^*(u)\mathbf{a}_i^*(v)d(i)^{-1}}{\|\hat{\mathbf{a}}^*(u)\|\,\|\hat{\mathbf{a}}^*(v)\|}\,, \end{aligned} \tag{8}$$

with $\langle \mathbf{x}, \mathbf{y}\rangle$ as the inner product between vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. The more nodes are similarly activated in both spreading processes, one starting at node u and one at v, the more similar u and v are. This measure allows for a detection of dense communities and requires a direct and indirect neighborhood overlap, as can be seen in Figure 1b. Node 1 is more similar to $\{4, 5, 6, 7\}$ than to 2 or 3 even if 1 is automorphically equivalent to 2 and 3. In [20] this kind of node similarity is categorized as closeness similarity.

The computation of node similarities proposed in [19] can be seen in terms of spreading activation as well. The accumulated and normalized activation values themselves represent the similarities between the activated nodes and the node at which the spreading process started. As stated, their method is applicable only on undirected graphs. For directed graphs the activation values are not necessarily symmetric, yielding asymmetric similarities.

5.2 Signature Similarity

The second similarity is based on the comparison of the amount of activation changes during spreading activation processes and named *signature similarity*. For each node a signature vector can be determined, consisting of velocity vector norms (see Section 4). The direction of the velocity vectors represent the change of direction of the activation patterns and their norms represent the step size between subsequent iterations towards the principal eigenvector of the weight matrix W. By the comparison of the signature

vectors a structural similarity can be derived. In this work we use the cosine measure to compare the signature vectors, thus the signature similarity is denoted as

$$\begin{aligned}\sigma_{\text{sig}}(u,v) &= \cos(\tau(u),\tau(v)) \\ &= \frac{\langle \tau(u),\tau(v)\rangle}{\|\tau(u)\|\,\|\tau(v)\|} \\ &= \frac{\sum_{k=1}^{k_{max}} \left\|\delta^{(k)}(u)\right\| \left\|\delta^{(k)}(v)\right\|}{\|\tau(u)\|\,\|\tau(v)\|}\,.\end{aligned} \tag{9}$$

Nodes that are similar due to the activation similarity have to be close to each other in the graph, since the same direct and indirect neighbor nodes need to be activated similarly. The signature similarity is not based on the activation pattern itself but on the amount of change of these patterns. If the structure of the neighborhood of two nodes is similar, the change of activation will be similar too, and thus the signature similarity will yield higher values as if the structure is different.

A similar step size between two subsequent iterations yields from a similar structure, i.e. the nodes $\{1, 2, 3\}$ (black) of Figure 1a are not distinguishable by their signature vectors, since they are automorphic images from each other. Whereas the activation vectors of these nodes are different, as well as the corresponding velocity vectors, the amount of change of direction of the activation vectors in each iteration is equal. Nodes do not necessarily have to be located in the same densely connected region to have a high signature similarity. This makes the signature similarity not a closeness but a structural similarity measure. Nodes with a structurally similar neighborhood have a high signature similarity even if they are located far apart from each other. An overlapping neighborhood is thereby not necessary, which can be seen in Figure 1a, where all the leaf nodes (white) have a signature similarity value of 1, even if their direct neighborhood is not overlapping at all.

6 Experiments

To demonstrate our approach we apply the two kinds of node similarities to the Schools-Wikipedia[2] (2008/09) dataset. The first aim is to find result nodes that are structurally similar to given query nodes by using the signature similarity. Secondly we want to find nodes that are closely connected (directly or indirectly) to the query nodes or interesting result nodes, respectively, using the activation similarity, and extract the corresponding subgraphs. Since the extraction of communities is not the aim of this work we do not focus on this issue. Instead we consider the induced subgraph of the k most similar nodes based on the activation similarity according to a query node, as dense local neighborhood, or community of that query.

Once structurally similar nodes have been detected and the corresponding communities have been extracted and illustrated by means of centrality layouts, we manually compare these subgraphs in order to find structural coherences. We are thereby interested in the status or rank of the result nodes in their community and the most central

[2] `http://schools-wikipedia.org/`

nodes. Our assumption is that the communities of the result nodes are similar, in these terms, to the community of the query node.

6.1 Schools-Wikipedia

The Schools-Wikipedia (2008/09) dataset consists of a subset of the English Wikipedia[3] dataset, with around 5500 articles. The articles are grouped into 154 different categories, consisting of 16 main or top-level categories, where each article is assigned to at least one category. As in Wikipedia, articles can reference other articles via hyperlinks. In Schools-Wikipedia external links have been filtered.

To create the graph, each article is considered as a unit of information and modeled as a node. Each hyperlink that connects articles is considered as a relation between two units of information and represented as an undirected edge connecting the corresponding nodes. The resulting graph consists of four connected components, whereas three of the components consist only of one node and are also filtered. Convergence of all spreading activation processes on the filtered graph is ensured by connectedness, non-bipartiteness and undirectedness. Table 1 lists some basic properties of the remaining graph.

Table 1. Basic graph properties of the filtered Schools-Wikipedia graph

Schools-Wikipedia graph properties	
Number of nodes	5536
Number of edges	190149
Minimal node degree	1
Maximal node degree	2069
Average node degree	68.7
Diameter	5

We applied spreading activation processes as described in Section 3 to the graph, in order to compute the activation and signature similarities between all nodes, defined in Section 5. Since the spreading activation processes converge quickly due to the underlying graph structure, indicated e.g. by the small diameter, we only computed the first 10 iterations of each spreading process to compute the similarities. Concerning the activation similarity we used a decay value of $\alpha = 0.3$ to compute the accumulated activation vectors in order to focus on the local neighborhood of nodes. The choice of parameters is not discussed in this work. Here it is sufficient to mention that further iterations (> 10) do not contribute significantly to both similarities due to the small decay as well as the small diameter and thus fast convergence.

In our experiment we wanted to find well-known, scholarly persons from different areas of research, which play similar roles in their communities. Our focus is on well-known people, since the results can be reasonably evaluated based on general knowledge. The query consists of the node of the well-known Italian physicist *Galileo Galilei*.

[3] `http://en.wikipedia.org/wiki/Main_Page`

To find the structurally most similar persons, all nodes, which are assigned to the *People* category are sorted based on their corresponding signature similarity to the query. Since we focused only on people with a similar structural position, we filtered out all nodes not belonging to the People category. Additionally we were interested in the nodes belonging to the community around Galileo Galilei. Therefore we considered all nodes, not only those assigned to the People category and sorted them according to their activation similarity to the query. Table 2 lists the 10 most similar nodes, as well as the 16th and 17th nodes of the People category, based on the signature similarity, and the 10 most similar nodes, as well as the 16th and 17th nodes of all categories based on the activation similarity, compared to Galileo Galilei.

Table 2. 10 most similar nodes to Galileo Galilei and the 16th and 17th nodes; *left* assigned to the *People* category, based on the signature similarity; *right* of all categories, based on the activation similarity.

Galileo Galilei		
Rank	Signature similarity	Activation Similarity
1	Galileo Galilei	Galileo Galilei
2	Isaac Newton	Johannes Kepler
3	Johannes Kepler	Heliocentrism
4	Aristotle	Nicolaus Copernicus
5	Leonhard Euler	Isaac Newton
6	Mary II of England	Phil. Nat. Principa Mathematica
7	James Clerk Maxwell	Kepler's laws of planetary motion
8	Anne of Great Britain	Classical mechanics
9	James I of England	History of physics
10	Henry VII of England	Astronomy
⋮	⋮	⋮
16	Plato	Newton's laws of motion
17	Euclid	General relativity

It can be seen that Galileo himself is the most similar node, which makes sense in terms of the cosine similarity used on activation and signature vectors. Nodes such as *Heliocentrism*, *Astronomy*, *History of physics*, etc. are part of his closer community, reasonably, since he worked primarily in these fields and played a major role in them. Among others Galileo is called "the father of modern physics". Other important scientists who played a major role in these areas as well, such as *Nicolaus Copernicus*, *Johannes Kepler*, and *Isaac Newton* are also part of his community.

On inspecting the structurally similar nodes, the names *Plato* and *Euclid* attract our attention; they are the 16th and 17th structurally most similar nodes of the People category. Both men played a major role in their areas of research too, philosophy and mathematics, respectively, which are different to those of Galileo. Plato contributed significantly to the foundations of Western philosophy and Euclid is said to be the "father of geometry". Newton and Kepler, have a high signature similarity as well and played - like Galilei - a major role in their areas of research too. However, their areas of

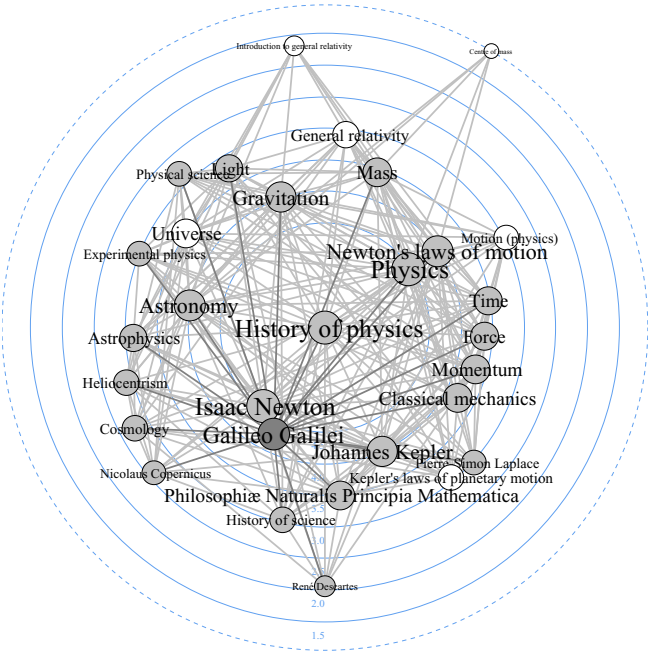
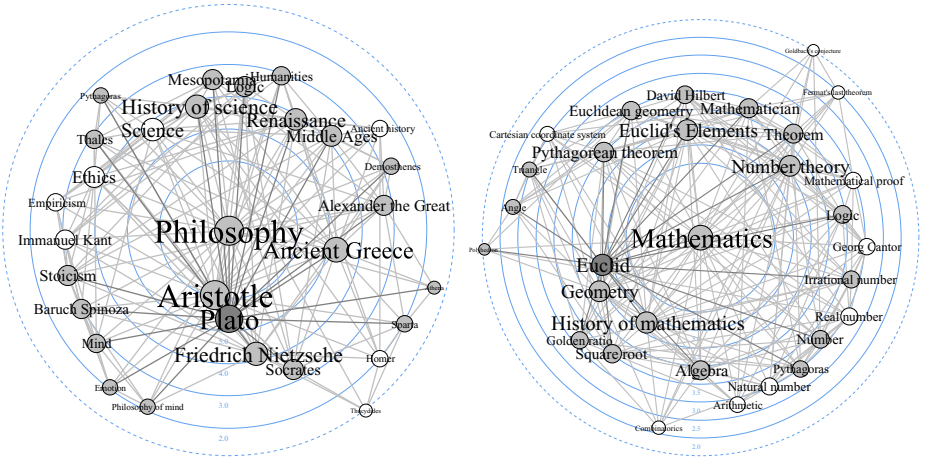


Fig. 3. The subgraph of the 30 most similar nodes to Galileo Galilei based on activation similarity. The used layout is a centrality layout, based on eigenvector centrality.

research do not differ to Galilei's as much as those of Plato and Euclid do. In terms of structural bisociations, structurally similar nodes, that represent units of information in unrelated fields of knowledge are potentially more interesting, than those in the same or similar fields. As a result of this fact and due to their high degree of popularity, Plato and Euclid were chosen in order to compare their communities.

We extracted the induced subgraphs of their communities consisting of the 30 most similar nodes based on the corresponding activation similarities. We used 30 nodes, since the subgraphs of this size can be visualized in a reasonable manner and both structural similarities as well as differences can be shown. Figure 3 shows the community around Galileo, Figure 4a that around Plato and Figure 4b that around Euclid.

Nodes are represented as circles, whereas the corresponding size and the size of the label is proportional to their degree. The nodes of the corresponding persons are emphasized by dark gray; their direct neighbors are light gray and all other nodes white. The layout of all graphs is a centrality layout based on eigenvector centrality [7]. The eigenvector centrality is, like other centrality indices, a measure to quantify the status of nodes in a graph. The higher the value compared to others, the more central or important the node, and the more central its position in the visualization. In contrast, the lower the value, the lower its status or importance and the more peripherical the position.



(a) (b)

Fig. 4. Two subgraphs of the 30 most similar nodes to Plato (4a) and Euclid (4b), based on the activation similarity. The used layout is a centrality layout, based on the eigenvector centrality.

It can be seen that Galileo, Plato, and Euclid are connected to most of the nodes of their communities. Galileo has 23 direct neighbors, Plato 22 and Euclid 21. Their neighborhoods can roughly be partitioned into three semantic groups: the fields they worked in, topics and issues important in these particular fields and other important persons who contributed significantly to these fields as well. In the case of Galileo, the fields of research are *Physics*, *Astronomy*, *Classical mechanics* etc. Important topics and issues in these fields are e.g. *Gravitation*, *Mass*, and *Force* and other important persons who worked in these fields are e.g. *Nicolaus Copernicus*, *Isaac Newton* or *Johannes Kepler*. In the case of Plato, the fields of research are *Philosophy* and *Philosophy of mind*. Important topics are e.g. *Emotion* and *Logic* and other important persons are *Aristotle* and *Socrates*. In the case of Euclid, the fields of research are *Mathematics* and *Euclidean geometry*. Important issues are e.g. *Angle* and *Triangle* and other important persons are *Pythagoras* and *David Hilbert*. Even if Galileo, Plato, and Euclid are directly connected to most of the nodes in their community, the most central nodes are, however the fields for which (among others) they are famous for: *History of physics*, *Philosophy*, and *Mathematics*. Nevertheless their status is very central compared to all other nodes in the corresponding communities.

In all three subgraphs there exist other nodes with a similar centrality. In Galiliei's community these nodes are *Isaac Newton*, *Johannes Kepler*, *Physics*, *Astronomy*, and *Gravitation*, in Plato's *Aristotle*, and *Anchient Greece* and in Euclid's *Geometry*, *Euclids Elements*, and *History of mathematics*. All of these nodes, except *Euclids Elements* and *Gravitation* have a high signature similarity according to Galilei, even though some nodes, such as *Aristotle* are not part of his community and thus do not have a high degree of activation similarity. However, the signature similarity of *Euclids Elements* and *Gravitation* is also not very low. The nodes are part of the 270 most similar nodes of all categories. Additionally it can be seen that in the case of Galileo and Plato, there

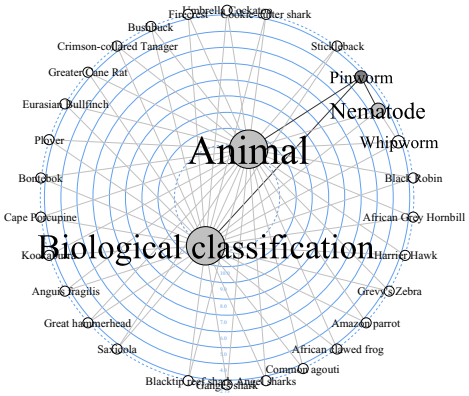


Fig. 5. The subgraph of the 30 most similar nodes to Pinworm, based on the activation similarity. The used layout is a centrality layout, based on the eigenvector centrality.

is one slightly more central node in the People category: *Isaac Newton* and *Aristotle*, respectively. In all three communities the most peripheric nodes are not connected to Galileo, Plato and Euclid.

In a nutshell, structural coherences of nodes with high signature similarity and their corresponding communities can be seen based on various aspects, such as their own status or centrality and those of other nodes, their connectedness, their degree as well as the density of their community, etc.

On the one hand very similar nodes to a certain query, based on signature similarity, are interesting to show structural coherences. On the other, very dissimilar nodes are also interesting to show structural differences. The most dissimilar node to Galileo is *Pinworm*. Again we extracted the 30 nodes most similar to Pinworm, based on activation similarity, and illustrated the induced subgraph in Figure 5 using the centrality layout.

Nodes are represented as circles, whereas size and labelsize are proportional to their degree, up to a certain maximum size. The Pinworm node is emphasized by dark gray, its direct neighbors are light gray and all other nodes white. The structural differences of the Pinworm node as well as its community can be seen clearly. Pinworm has, as well as almost all other nodes a very peripheric position. Additionally the most central position is shared by the two nodes *Animal* and *Biological classification*, which are connected to all nodes of the community. In addition the density of the community is much lower then those of the communities of Galilei, Plato, or Euclid.

7 Conclusion

In this work we have shown how two kinds of similarities to compare nodes in a graph can be derived from spreading activation processes. The activation similarity is based on the comparison of accumulated activation vectors and yields a spatial or closeness similarity. The signature similarity is based on the comparison of norms of velocity

vectors and yields a structural similarity. By applying both kinds of similarities we can find structurally similar nodes on the one hand, which are not necessarily located close together, and dense subgraphs or communities around nodes on the other. We applied this procedure to the Schools-Wikipedia (2008/09) dataset and preliminary results are very encouraging: the nodes of Euclid, and Plato for example, are structurally similar to that of Galileo Galilei. By comparing their communities structural similarity was able to be confirmed manually. The experiments suggested that the combination of these two kinds of similarities is a promising tool in terms of identification and extraction of structural bisociations.

References

1. Anderson, J.R.: A spreading activation theory of memory. Journal of Verbal Learning and Verbal Behavior 22(3), 261–295 (1983)
2. Anderson, J.R., Pirolli, P.L.: Spread of activation. Journal of Experimental Psychology: Learning, Memory, and Cognition 10(4), 791–798 (1984)
3. Aswath, D., Ahmed, S.T., D'cunha, J., Davulcu, H.: Boosting item keyword search with spreading activation. In: Web Intelligence, pp. 704–707 (2005)
4. Belew, R.K.: Adaptive information retrieval: using a connectionist representation to retrieve and learn about documents. In: SIGIR 1989: Proceedings of the 12th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 11–20. ACM Press, New York (1989)
5. Berthold, M.R., Brandes, U., Kötter, T., Mader, M., Nagel, U., Thiel, K.: Pure spreading activation is pointless. In: Proceedings of the CIKM the 18th Conference on Information and Knowledge Management, pp. 1915–1919 (2009)
6. Berthold, M.R., Dill, F., Kötter, T., Thiel, K.: Supporting Creativity: Towards Associative Discovery of New Insights. In: Washio, T., Suzuki, E., Ting, K.M., Inokuchi, A. (eds.) PAKDD 2008. LNCS (LNAI), vol. 5012, pp. 14–25. Springer, Heidelberg (2008)
7. Bonacich, P.: Factoring and weighting approaches to status scores and clique identification. Journal of Mathematical Sociology 2, 113–120 (1972)
8. Brandes, U., Erlebach, T.: Network Analysis: Methodological Foundations. Springer (2005)
9. Collins, A.M., Loftus, E.F.: A spreading-activation theory of semantic processing. Psychological Review 82(6), 407–428 (1975)
10. Crestani, F.: Application of spreading activation techniques in information retrieval. Artif. Intell. Rev. 11(6), 453–482 (1997)
11. Dubitzky, W., Kötter, T., Schmidt, O., Berthold, M.R.: Towards Creative Information Exploration Based on Koestler's Concept of Bisociation. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 11–32. Springer, Heidelberg (2012)
12. Duch, W., Matykiewicz, P., Pestian, J.: Neurolinguistic approach to natural language processing with applications to medical text analysis. Neural Networks 21, 1500–1510 (2008)
13. Hopfield, J.J.: Neurons with graded response have collective computational properties like those of two-state neurons. Proceedings of the National Academy of Sciences 81(10), 3088–3092 (1984)
14. Jaccard, P.: Étude comparative de la distribution florale dans une portion des alpes et des jura. Bulletin del la Société Vaudoise des Sciences Naturelles 37, 547–579 (1901)

15. Katz, L.: A new status index derived from sociometric analysis. Psychometrika 18(1), 39–43 (1953)
16. Kosinov, S., Marchand-Maillet, S., Kozintsev, I.: Dual diffusion model of spreading activation for content-based image retrieval. In: MIR 2006: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 43–50. ACM, New York (2006)
17. Kötter, T., Berthold, M.R.: From Information Networks to Bisociative Information Networks. In: Berthold, M.R. (ed.) Bisociative Knowledge Discovery. LNCS (LNAI), vol. 7250, pp. 33–50. Springer, Heidelberg (2012)
18. Kötter, T., Thiel, K., Berthold, M.R.: Domain bridging associations support creativity. In: Proceedings of the International Conference on Computational Creativity, pp. 200–204 (2010)
19. Leicht, E.A., Holme, P., Newman, M.E.J.: Vertex similarity in networks. Physical Review E 73(2), 026120 (2006)
20. Lerner, J.: Structural Similarity of Vertices in Networks. PhD thesis, Universität Konstanz, Universitätsstr. 10, 78457 Konstanz (2007)
21. Liu, W., Weichselbraun, A., Scharl, A., Chang, E.: Semi-automatic ontology extension using spreading activation. Journal of Universal Knowledge Management, 50–58 (2005)
22. Moody, J.: Peer influence groups: identifying dense clusters in large networks. Social Networks 23(4), 261–283 (2001)
23. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications 10(2), 191–218 (2006)
24. Quillian, M.R.: A revised design for an understanding machine. Mechanical Translation 7, 17–29 (1962)
25. Quillian, M.R.: Semantic memory. In: Minsky, M. (ed.) Semantic Information Processing, pp. 227–270. The MIT Press, Cambridge (1968)
26. Salton, G., Buckley, C.: On the use of spreading activation methods in automatic information retrieval. Technical report, Dept. Computer Science, Cornell Univ., Ithaca, NY (1988)
27. Salton, G.: Automatic Information Organization and Retrieval. McGraw Hill (1968)
28. Ziegler, C.N., Lausen, G.: Spreading activation models for trust propagation. In: 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service, EEE 2004, pp. 83–97. IEEE Computer Society, Washington, DC (2004)