Bisociative Discovery of Interesting Relations Between Domains

Uwe Nagel, Kilian Thiel, Tobias Kötter, Dawid Piątek, and Michael R. Berthold

Nycomed-Chair for Bioinformatics and Information Mining
Dept. of Computer and Information Science
University of Konstanz
firstname.lastname@uni-konstanz.de

Abstract. The discovery of surprising relations in large, heterogeneous information repositories is gaining increasing importance in real world data analysis. If these repositories come from diverse origins, forming different domains, domain bridging associations between otherwise weakly connected domains can provide insights into the data that can otherwise not be accomplished. In this paper, we propose a first formalization for the detection of such potentially interesting, domain-crossing relations based purely on structural properties of a relational knowledge description.

1 Motivation

Classical data mining approaches propose two major alternatives to make sense of knowledge representing data collections. One is to formulate specific, semantic queries on the given data. However, this is not always useful since users often do not know ahead of time what exactly they are searching for. Alternatively, Explorative (or Visual) Data Mining attempts to overcome this problem by creating a more abstract overview of the entire data together with subsequent drill-down operations. Thereby it additionally enables the search for interesting patterns on a structural level, detached from the represented semantical information. However, such overviews still leave the entire search for interesting patterns to the user and therefore often fail to actually point to interesting and truly novel details.

In this paper we propose an approach to explore integrated data by finding unexpected and potentially interesting connections that hopefully trigger the user's interest, ultimately supporting creativity and outside-the-box thinking. The approach we propose attempts to find such unexpected relations between seemingly unrelated domains. As pointed out by Henri Poincaré [11]: "Among chosen combinations the most fertile will often be those formed of elements drawn from domains which are far apart... Most combinations so formed would be entirely sterile; but certain among them, very rare, are the most fruitful of all." Consequently, instead of only fusing different domains and sources to gain

a large knowledge base, we try to identify (possibly hidden) domains and search for rare instead of frequent patterns, i.e. exclusive, domain crossing connections.

In this paper we assume a knowledge representation fulfilling very few conditions and address two subproblems: the identification of domains and the assessment of the potential interestingness of connections between these domains.

2 Networks, Domains and Bisociations

In this section, we transfer the theoretical concept of domain crossing associations which are called *bisociations* [8] (to emphasize the difference to associations within a single domain) to a setting where a relational description of knowledge is given. We will explain the model that incorporates our knowledge base, narrow down the basic theoretical concepts underlying *domains* and *bisociations* and identify corresponding measures of interestingness.

2.1 Knowledge Modeling

As a preliminary, we assume that the available knowledge is integrated into a unifying data model. We model this as an undirected, unweighted graph structure with nodes representing units of information and edges representing their relations. Examples for information units are terms, documents, genes or experiments. Relations could arise from references, co-occurrences or explicitly encoded expert knowledge. A graph is described as $G = (V, E)$ with node set V, edge set $E \subseteq \binom{V}{2}$ and $n = |V|$ the number of nodes. The degree of a node, i.e. the number of incident edges, is denoted as $d(v)$ and we access the structure of G via its adjacency matrix A, with $(A)_{uv} = 1$ if $\{u, v\} \in E$ and 0 otherwise.

An important aspect of the model is the semantic of the employed links. We consider two different types of links, which either express similarity or another semantic relation. Consider for example a knowledge network with scientific articles as information units. Links in a derived network could either encode similarities between articles or the fact that one article references another.

2.2 Domains

In this context, a *domain* is a set of information units from the same field or area of knowledge. Domains exist with different granularity and thus can be (partially) ordered in a hierarchical way from specific to general. An example is provided by the domains of quantum physics, physics in general, and science. Consequently, the granularity of a domain depends on a specific point of view, which can be a very local one.

Due to their hierarchical nature, information units can belong to several domains which are not necessarily related. E.g. the eagle belongs to the animal domain and in addition to the unrelated coat of arms domain.

Intuitively, a set of highly interconnected nodes indicates an intense interrelation that should be interpreted as a common domain. While this is a sound

assumption when connections express similarities between the involved concepts, it is not true when links express other semantic relations. Consider for example scientific articles approaching a common problem. The similarity of these articles is not necessarily reflected by mutual references, especially if they were written at the same time. However, they will very likely share a number of references. Consequently, we derive domains from common neighborhoods instead of relying on direct connections between information units. This allows domains to be identified when the connections express either references or similarities since densely connected nodes also have similar neighborhoods.

Domain Recovery Two information units that share all (or - more realistically - almost all) their connections to other information units should therefore belong to a common domain. Since they are in this respect indistinguishable and their relations form the sole basis for our reasoning about them, all possibly identifiable domains have to contain either both or none of them. We will show a node similarity that expresses this property and relaxes the conditions. Recursive merging of nodes based on this similarity leads to a merge tree as produced by hierarchical clustering. Consequently, we consider the inner nodes of this merge tree as candidates for domains. Note that this clustering process is distinguished from classical graph clustering by the employed node similarity.

The resulting domains form a hierarchy on the information units which is similar to an ontology. I.e. considering two arbitrary domains, either one domain is completely contained in the other, or they are disjoint. Apparently, a number of domains could remain unidentified since the set of domains is not restricted to hierarchies but could also contain partially overlapping domains. We consider this as an unavoidable approximation for now, posing the extraction of domains as a separate problem.

2.3 Bisociations

A connection - usually indirect - between information units from multiple, otherwise unrelated domains is called *bisociation* in contrast to associations that connect information units within the same domain. The term was introduced by Koestler [7] in a theory to describe the creative act in humor, science and art. An example of a creative discovery triggered by a bisociation is the theory of electromagnetism by Maxwell [9] that connects electricity and magnetism.

Up to now, three different patterns of bisociation have been described in this context: bridging concepts, bridging graphs and structural similarity [8]. Here we focus on the discovery of bridging graphs, i.e. a collection of information units and connections providing a "bisociative" relation between diverse domains.

Among the arbitrary bisociations one might find, not all are going to be interesting. To assess their interestingness, we follow Boden [2] defining a creative idea in general as *new*, *surprising*, and *valuable*. All three criteria depend on a specific reference point: A connection between two domains might be long known to some specialists but new, surprising, and hopefully valuable to a specific observer, who is not as familiar with the topic. To account for this, Boden [2]

defines two types of creativity namely H-creativity and P-creativity. While H-creativity describes globally (historical) new ideas, P-creativity (psychological) limits the demand of novelty to a specific observer. Our findings are most likely to be P-creative since the found connections have to be indicated by the analyzed data in advance. However a novel combination of information sources could even lead to H-creative bisociations. Analog to novelty, the value of identified bisociations is a semantically determined property and strongly depends on the viewers' perspective. Since both novelty and value cannot be judged automatically, we leave their evaluation to the observer. In contrast, the potential surprise of a bisociation can be interpreted as the unlikeliness of a connection between the corresponding domains. We will express this intuition in more formal terms and use it as a guideline for an initial evaluation of possible bisociations.

Identifying Bisociations. Based on these considerations, we now characterize the cases where a connection between two domains forms a bisociation. In the graph representation, two domains are connected either directly by edges between their nodes or more generally by nodes that are connected to both domains - the *bridging nodes*. These connecting nodes or edges bridge the two domains and together with the connected domains they form a *bisociation candidate*:

Definition 1 (Bisociation Candidate). *A* bisociation candidate *is a set of two domains and their connection within the network.*

Since it is impossible to precisely define what a surprising bisociation is, we rather define properties that distinguish promising bisociation candidates: *exclusiveness*, *size*, and *balance*. These can be seen as technical demands derived from a more information-scientific view as e.g. expressed in [5]: In Ford's view, the creativity of a connection between two domains is related to (i) the dissimilarity of the connected domains and (ii) the level of abstraction on which the connection is established. In the following we try to transport these notions into graph theoretic terms by capturing them in technical definitions. Therein we interpret the dissimilarity of two domains as their mutual reachability by edges restricted to short connections: either direct by edges linking nodes of the different domains or indirect by nodes connected to both domains. Thus dissimilarity relates to the exclusiveness of the bisociation candidate: maximal dissimilarity is obviously rendered by two completely unconnected domains, closely followed by "minimally connected" domains. While the former case obviously does not yield a bridging graph based bisociation (i.e. the connection itself is missing) the latter is captured by exclusiveness.

Exclusiveness states that a bisociation is a rare connection between the two domains, rendering the fact that bisociations are surprising connections between dissimilar domains. At the same time it excludes local exclusivity caused by nodes of high degree which connect almost everything, even unrelated domains, without providing meaningful connections.

Definition 2 (Exclusiveness). *A bisociation candidate is exclusive iff its domains are bridged by a graph that is small in relation to the domains and which provides only few connections that are focused on the two domains.*

This can additionally be related to connection probabilities: consider the probability that two nodes from diverse domains are related by a direct link or an intermediate node. If only a few of these relations exist between the domains, the probability that such a pair of randomly chosen information units is connected is low and thus the surprise or unlikeliness is high.

Directly entangled with this argument is the demand for size: a connection consisting of only a few nodes and links becomes less probable with growing domain sizes. In addition, a relation between two very small domains is hard to judge. It could be an expression of their close relation being exclusive only due to the small size of the connected domains. In that case the larger domains containing these two would show even more relations. It could also be an exclusive link due to domain dissimilarity. However, this situation would in turn be revealed when considering the larger domains, since these would also be exclusively connected. In essence, the exclusiveness of such a connection is pointless if the connected domains are very small, while it is amplified by domains of larger size. We formalize this in the following definition:

Definition 3 (Size). *The size of a bisociation candidate is the number of nodes in the connected domains.*

In terms of [5] the demand for size relates to the level of abstraction. Obviously a domain is more abstract than its subdomains and thus an exclusive link between larger (i.e. more abstract) domains is a more promising bisociation than a link between smaller domains.

Finally, the balance property assures that we avoid the situation of a very small domain attached to a large one:

Definition 4 (Balance). *A bisociation candidate is balanced iff the connected domains are of similar size.*

In addition, domains of similar size tend to be of similar granularity and are thus likely to be on comparable levels of abstraction. Thereby the demand for balance avoids exclusive links to small subdomains that are actually part of a broader connection between larger ones.

Summarizing, a bisociation candidate is promising if it is exclusive, of reasonable size, and balanced.

3 Finding and Assessing Bisociations

In this section, we translate the demands described in Section 2 into an algorithm for the extraction and rating of bisociations. Therein we follow the previously indicated division of tasks: (i) domain extraction and (ii) scoring of bisociation candidates.

3.1 Domain extraction

As described in Section 2, domain affiliation of nodes is reflected by similar direct and indirect neighborhoods in the graph. Thus comparing and grouping nodes

based on their neighborhoods yields domains. In the following, we establish the close relation of a node similarity measure called *activation similarity* [12] to the above described demands. Based on this similarity, we show in a second part how domains can be found using hierarchical clustering.

Activation similarity The employed node similarity is based on *spreading activation* processes in which initially one node is activated. The activation spreads iteratively from the activated node, along incident edges, to adjacent nodes and activates them to a certain degree as well. Given that the graph is connected and not bipartite, the process converges after sufficient iterations. The final activation states are determined by the principal eigenvector of the adjacency matrix of the underlying graph as shown in [1]. Activation states of all nodes at a certain time k are represented by the activation vector $\mathbf{a}^{(k)} \in \mathbb{R}^n$ defined by $\mathbf{a}^{(k)} = A^k\mathbf{a}^{(0)} / \left\|A^k\mathbf{a}^{(0)}\right\|_2$, where the value $\mathbf{a}_v^{(k)}$ ($\mathbf{a}^{(k)}$ at index v) is the activation level of node $v \in V$. Then $\mathbf{a}_v^{(k)}(u)$ represents the activation of node v at time k, induced by a spreading activation process started at node u, i.e. with $\mathbf{a}_u^{(0)} = 1$ and $\mathbf{a}_v^{(0)} = 0$ for $v \neq u$. This reflects the relative (due to normalization) reachability of node v from node u via walks of length k. More precisely, it represents the weighted fraction of weighted walks of length k from u to v among all walks of length k started at u. In order to consider more than just walks of a certain length, the activation vectors are normalized and accumulated with an additional decay $\alpha \in [0, 1)$ to decrease the impact of longer walks. The *accumulated activation vector* of node u is then defined by $\hat{\mathbf{a}}^*(u) = D^{-\frac{1}{2}} \left(\sum_{k=1}^{k_{\max}} \alpha^k \mathbf{a}^{(k)}(u)\right)$, with $D = \mathrm{diag}(d(v_1), \ldots, d(v_n))$ being the degree matrix and $k_{\max}$ the number of spreading iterations. The degree normalization is useful to account for nodes of a very high degree. These are more likely to be reached and would thus distort similarities if not taken care of. The value $\hat{\mathbf{a}}_v^*(u)$ represents the (normalized) sum of weighted walks of different lengths $1 \leqslant k \leqslant k_{max}$ from u to v proportional to all weighted walks of different length starting at u and thus the relative reachability from u to v.

In essence, the vector $\hat{\mathbf{a}}^*(v)$ describes the reachability of other nodes from v and thereby its generalized neighborhood. On this basis, we use the activation similarity $\sigma_{\mathrm{act}}(u, v) = \cos(\hat{\mathbf{a}}^*(u), \hat{\mathbf{a}}^*(v))$ of nodes u and v to compare their neighborhoods. In case of identical neighborhoods, activation spreads identically, resulting in a similarity of 1. If the same nodes can be reached similarly from u and v the similarity between them is high, which corresponds with our assumption about the properties of domains. For usual reasons, we will use the corresponding distance $1 - \sigma_{\mathrm{act}}(u, v)$ for hierarchical clustering.

Domain identification We employ hierarchical clustering for domain identification using Ward's linkage method [13], which minimizes the sum of squared distances within a cluster. This tends to produce compact clusters and to merge clusters of similar size and thus corresponds well with the notion of a domain.

First of all, we would expect a certain amount of similarity for arbitrary information units within a domain and thus a compact shape. Further, clusters of similar size are likely to represent domains on the same level of granularity and thus merging those corresponds to building upper-level domains. The resulting *merge tree* is defined as follows:

Definition 5 (Merge tree). *A merge tree $T = (V_T, E_T)$ for a graph $G = (V, E)$ is a tree produced by a hierarchical clustering with node set $V_T = V \cup \Lambda$ where Λ is the set of clusters obtained by merging two nodes, a node and a cluster or two clusters. E_T describes the merging structure: $\{u\lambda, v\lambda\} \subseteq E_T$ iff the nodes or clusters u and v are merged into cluster $\lambda \in \Lambda$.*

However, not all clusters in the hierarchy are good domain candidates. If a cluster is merged with a single node, the result is unlikely to be an upper-level domain. Most likely, it is just an expansion of an already identified domain resulting from agglomerative clustering. These considerations lead to the domain definition:

Definition 6 (Domain). *A cluster δ_1 is a domain iff in the corresponding merge tree it is merged with another cluster:*

$$\delta_1 \in \Lambda \text{ is a domain} \Leftrightarrow \exists \delta_2, \kappa \in \Lambda \text{ such that } \{\{\delta_1, \kappa\}, \{\delta_2, \kappa\}\} \subseteq E_T \ .$$

I.e. a cluster is a domain, if it is merged with another cluster.

3.2 Scoring bisociation candidates

In the next step, we iterate over all pairs of disjoint domains and construct a bisociation candidate for each pair by identifying their bridging nodes:

Definition 7 (Bridging nodes). *Let δ_1 and δ_2 be two domains derived from the merge tree of the graph $G = (V, E)$. A set of bridging nodes $\mathrm{bn}(\delta_1, \delta_2)$ is a set of nodes that are connected to both domains:*

$$\mathrm{bn}(\delta_1, \delta_2) = \{v \in V : \exists \{v, u_1\}, \{v, u_2\} \in E \text{ with } u_1 \in \delta_1, u_2 \in \delta_2\} \ .$$

Note that this definition includes nodes belonging to one of the two domains, thus allowing direct connections between nodes of these domains.

We now define the *b-score*, expressing the combination of exclusiveness, size, and balance as defined in Section 2. We therefore consider each property separately and combine them into an index at the end. Exclusiveness could be directly expressed by the number of nodes in $\mathrm{bn}(\delta_1, \delta_2)$. However, this is not a sufficient condition. Nodes of high degree are likely to connect different domains, maybe even some of them exclusively. Nevertheless, such nodes are unlikely to form good bisociations since they are not very specific. On the other hand, bridging nodes providing only a few connections at all (and thus a large fraction of them within δ_1 and δ_2) tend to express a very specific connection. Since we are only interested in the latter case, the natural way of measuring exclusiveness is by using the inverse of the sum of the bridging nodes' degrees: $2/\sum_{v \in \mathrm{bn}(\delta_1, \delta_2)} d(v)$.

The 2 in the numerator ensures that this quantity is bound to the interval $[0, 1]$, with 1 being the best possible value. The balance property is accounted for by relating the domain sizes in a fraction: $\min\{|\delta_1|, |\delta_2|\}/\max\{|\delta_1|, |\delta_2|\}$, again bound to $[0, 1]$ with one expressing perfect balance. Finally, the size property is integrated as the sum of the domain sizes.

As described above, a combination of all three properties is a necessary prerequisite for an interesting bisociation. Therefore, our bisociation score is a product of the individual quantities. Only in the case of $\mathrm{bn}(\delta_1, \delta_2) = \emptyset$ is our measure undefined. However, this situation is only possible if the domains are unconnected, so we define the score to be 0 in this case. For all non-trivial cases the score has strictly positive values and is defined as follows:

Definition 8 (b-score). *Let δ_1 and δ_2 be two domains, then the b-score of the corresponding bisociation candidate is*

$$b\text{-}score(\delta_1, \delta_2) = \frac{2}{\sum\limits_{v \in bn(\delta_1, \delta_2)} d(v)} \cdot \frac{\min\{|\delta_1|, |\delta_1|\}}{\max\{|\delta_1|, |\delta_2|\}} \cdot (|\delta_1| + |\delta_2|).$$

The above definition has two important properties. Firstly, it has an intuitive interpretation: In our opinion, an ideal bisociation is represented by two equally sized domains connected directly by a single edge or indirectly by a node connected to both domains. This optimizes the b-score, leaving the sum of the domain sizes as the only criterion for the assessment of this candidate. Further, every deviation from this ideal situation results in a deterioration of the b-score. Secondly, the calculation of the b-score only involves information about the two domains and their neighborhoods and not the whole graph, which is important when the underlying graph is very large.

3.3 Complexity and scalability

To compute the pairwise activation similarities, the accumulated activation vectors for all nodes need to be determined. This process is dominated by matrix vector multiplications yielding a complexity of $\mathcal{O}(n^3)$. Note however, that exploitation of the network sparsity and the quick convergence of the power iteration leads to a much more efficient calculation. The complexity of the overall process is dominated by the evaluation of bisociation candidates. Here, we propose to prune the set of candidates by removing small domains and filter highly unbalanced candidates. E.g. in the example of Section 4 roughly 75% of all bisocation candidates involved domains with less than 4 nodes.

4 Preliminary Evaluation

To demonstrate our approach, we applied our method to the Schools-Wikipedia (2008/09) dataset[1]. Following the described method, we evaluated every pair of

[1] For detailed description of the Schools-Wikipedia dataset see [12].

disjoint domains and manually explored the top rated bisociation candidates to verify the outcome of our method.

The dataset consists of a subset of the English Wikipedia with about 5500 articles. For our experiment, we consider each article as a separate unit of information and model it as a node. We interpret cross-references as relations and introduce an undirected edge whenever one article references another. The resulting graph is connected except for two isolated nodes which we removed beforehand.

For the remaining nodes we extracted the domains as described. To focus on the local neighborhood of nodes we used the decay value $\alpha = 0.3$. Due to this decay and the graph structure the activation processes converged quickly allowing a restriction to $k_{\max} = 10$ iterations for each process. This choice seems arbitrary, but we ensured that additional iterations do not contribute significantly to the distances. First of all, the values of the following iterations tend to vanish due to the exponentially decreasing scaling factor, e.g. 0.3^{-10} in the last iteration. In addition, the order of distances between node pairs is unchanged by the additional iterations. Altogether we extracted 4,154 nested domains resulting in 8,578,977 bisociation candidates.

A part of a dendrogram involving birds is shown in Figure 1 to illustrate that our clustering yields conceptually well defined domains. In the example, birds of prey such as hawk, falcon, eagle etc. end up in the same cluster with carnivorous birds such as e.g. vulture, and are finally combined with non-carnivorous birds to a larger cluster. This example further illustrates that the nodes of a good domain are not necessarily connected, as there are few connections within the sets of birds, and yet they share a number of external references.

Since the b-scores of the best bisociation candidates (Figure 2) decrease quickly, we considered only the top rated pairs. The bisociation candidate with the best b-score is shown in Figure 3a. One of its domains contains composers while the other incorporates operating systems. These two seemingly unrelated domains are connected by *Jet Set Willy*, a computer game with a title music adapted from the first movement of Beethoven's Moonlight Sonata and a level editor for *Microsoft Windows*. Except for the small domain sizes, *Jet Set Willy* meets all formulated demands. Following three variants of the *Jet Set Willy* bisociation, the next best candidate is shown in Figure 3b. *Nine Million Bicycles* is a song connecting a south-east Asia domain with an astronomy domain. While Beijing is mentioned within the song itself, the corresponding article discusses lyrical errors concerning its statements about astronomical facts. To us these relations were new and surprising, though one might argue their value.

An example of a bisociation with more than one bridging node is shown in Figure 3c. The substantially lower b-score of bisociations with more than one bridging node is a result of their lower exclusiveness. An example of a poor bisociation can be seen in Figure 3d. Clearly, this is neither balanced nor exclusive (countries have a very high degree in Schools-Wikipedia) while its size is comparable to the other described candidates.

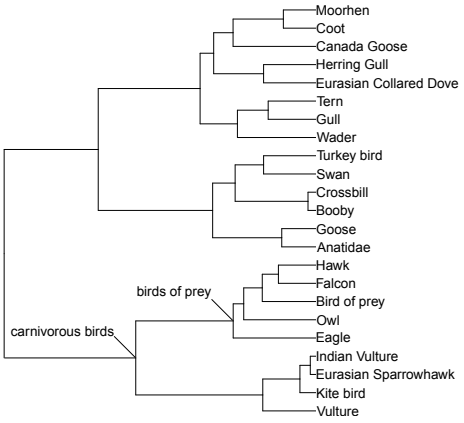


Fig. 1: Sub-dendrogram of articles about birds.

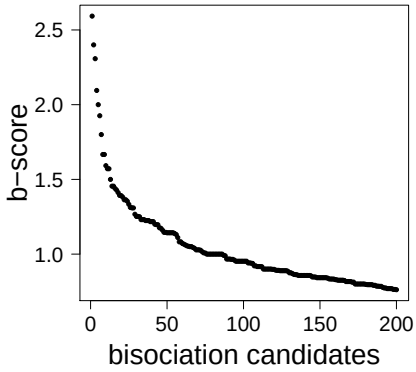
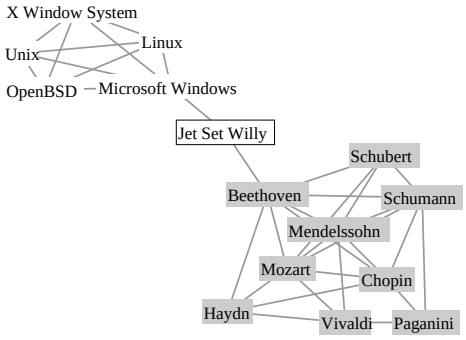
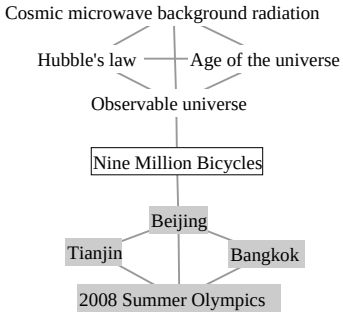


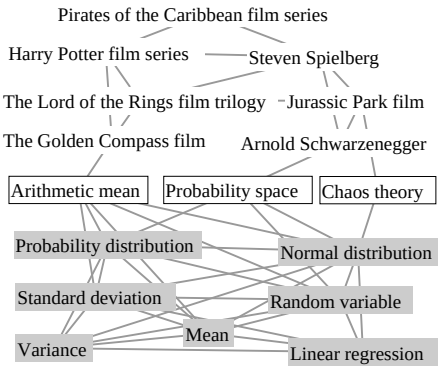
Fig. 2: Distribution of the b-score for the 200 top rated bisociation candidates.



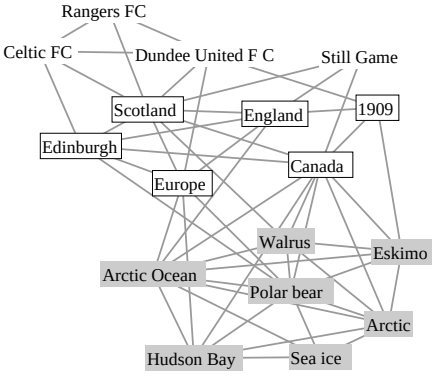
(a) b-score=2.59



(b) b-score=2.00



(c) b-score=0.35



(d) b-score=0.003

Fig. 3: Example bisociations and their b-score (see text for details).

The above examples illustrate that our index discriminates well with respect to exclusiveness and balance. A detailed examination showed in addition that size is negatively correlated with both other index components. This and the limited size of the dataset could explain the small sizes of the best rated candidates.

Our preliminary evaluation indicates the potential of the presented method to detect bisociations based on the analysis of the graph structure. Even though Schools-Wikipedia is a reasonable dataset for evaluation purposes, one cannot expect to find valuable or even truly surprising bisociations therein since it is limited to handpicked, carefully administrated common knowledge, suitable for children. We opted to manually evaluate the results since the value of a bisociation is a semantic property and highly subjective, inhibiting an automatic evaluation - although an evaluation on a dataset with manually tagged bisociations would be possible, if such a dataset were available. An evaluation using synthetic data is complicated by the difficulty of realistic simulation and could in addition introduce an unwanted bias on certain types of networks, distorting the results.

5 Related Work

Although a wealth of techniques solving different graph mining problems already exist (see e.g. [4] for an overview), we found none to be suitable for the problem addressed here. Most of them focus on finding frequent subgraphs, which is not of concern here. Closely related to our problem are clustering and the identification of dense substructures, since they identify structurally described parts of the graph. Yet bisociations are more complicated structures due to a different motivation and therefore require a different approach to be detected.

The exclusiveness of a connection between different groups is also of concern in the analysis of social networks. Especially *structural holes* and the notion of *betweenness* seem to address similar problems at first glance. Burt [3] regards the exlusiveness of connections in a network of business contacts as part of the capital a player brings to the competitive arena. He terms such a situation a *structural hole* that is bridged by the player. However, in his index only the very local view of the player is integrated, ignoring the structure of the connected domains. Further, his index would implicitly render domains a product of only direct connections between concepts, whereas we showed earlier that a more specific concept of similarity is advisable. A global measure for the amount of control over connections between other players is provided by betweenness [6]. Analog to structural holes, this concept captures one important aspect while missing the rest and thus fails to capture the overall concept.

Serendipitous discoveries strongly overlap with the bisociation concept since the involved fortuitousness is often caused by the connection of dissimilar domains of knowledge. Different approaches (e.g. [10]) exist to integrate this concept in recommender systems. They differ from bisociation detection in that they are concentrating on users' interests and not domains in general and are thus designed for a different setting and a different notion of optimality.

However, none of the mentioned approaches provide a coherent, formal setting applicable to bisociation detection.

6 Conclusion

We presented an approach for the discovery of potentially interesting, domain crossing associations, so-called bisociations. For this purpose we developed a formal framework to describe potentially interesting bisociations and corresponding methods to identify domains and rank bisociations according to interestingness. Our evaluation on a well-understood benchmark data set has shown promising first results. We expect that the ability to point the user to potentially interesting, truly novel insights in data collections will play an increasingly important role in modern data analysis.

Acknowledgements This research was supported by the DFG under grant GRK 1042 (Research Training Group „Explorative Analysis and Visualization of Large Information Spaces“) and the European Commission in the 7th Framework Programme (FP7-ICT-2007-C FET-Open, contract no. BISON-211898).

References

1. M. R. Berthold, U. Brandes, T. Kötter, M. Mader, U. Nagel, and K. Thiel. Pure spreading activation is pointless. In *Proceedings of the CIKM the 18th Conference on Information and Knowledge Management*, pages 1915–1919, 2009.
2. M. A. Boden. Précis of the creative mind: Myths and mechanisms. *Behavioral and Brain Sciences*, 17(03):519–531, 1994.
3. R. S. Burt. *Structural holes: the social structure of competition.* Harvard University Press, 1992.
4. D. J. Cook and L. B. Holder. *Mining graph data.* Wiley-Interscience, 2007.
5. N. Ford. Information retrieval and creativity: Towards support for the original thinker. *Journal of Documentation*, 55(5):528–542, 1999.
6. L. C. Freeman. A set of measures of centrality based upon betweenness. *Sociometry*, 40:35–41, 1977.
7. A. Koestler. *The Act of Creation.* Macmillan, 1964.
8. T. Kötter, K. Thiel, and M. R. Berthold. Domain bridging associations support creativity. In *Proceedings of the International Conference on Computational Creativity, Lisbon*, pages 200–204, 2010.
9. J. C. Maxwell. A treatise on electricity and magnetism. *Nature*, 7:478–480, 1873.
10. K. Onuma, H. Tong, and Ch. Faloutsos. Tangent: a novel, 'surprise me', recommendation algorithm. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '09, pages 657–666, 2009.
11. Henri Poincaré. Mathematical creation. *Resonance*, 5(2):85–94, 2000.
12. K. Thiel and M. R. Berthold. Node similarities from spreading activation. In *Proceedings of the IEEE International Conference on Data Mining*, pages 1085–1090, 2010.
13. J. H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.