Hecking *et al. Applied Network Science* (2019) 4:1
https://doi.org/10.1007/s41109-018-0108-x

Applied Network Science

RESEARCH Open Access

Positional analysis in cross-media information diffusion networks

Tobias Hecking*, Laura Steinert, Victor H. Masias and H. Ulrich Hoppe

*Correspondence:
hecking@collide.info
University of Duisburg-Essen,
Department of Computer Science
and Applied Cognitive Science,
Duisburg, Germany

Abstract

This paper describes a network reduction technique to reveal possibly hidden relational patterns in information diffusion networks of interlinked content published across different types of online media. Topic specific content items such as tweets (Twitter), web pages, or versions of Wikipedia articles can reference each other through hyperlinks, revisions, or retweet relationships, and thus, constitute a network that reflects the dissemination of information on the web. Beyond focusing on the structural linking of content items alone, the temporal aspect of information diffusion is explicitly taken into account by modelling the edge weight between two interlinked items according to the difference in their publication times. Non-negative matrix factorisation (NMF) is applied to decompose the resulting networks into groups of nodes occupying similar positions, which means that they have similar abilities to spread or receive information to or from other nodes. This allows for an easier observation of the basic underlying structure of cross-media information diffusion networks and their main information pathways. The utility of the approach and differences to other techniques will be demonstrated along two application scenarios related to two popular news stories and their dissemination in online media in 2016.

Keywords: Information diffusion, Non-negative matrix factorisation, Positional analysis

Introduction

Studying the diffusion of infections, information items, and opinions in networks of interrelated entities is one of the main challenges in the intersection of computational social science, epidemiology, and knowledge management. With regard to social media, insights into the underlying mechanisms of information diffusion support a better understanding of the emergence of public opinions as well as the identification of potential information bottlenecks. Moreover, it can provide insights into how information can be spread quickly among a large set of recipients, for example in case of emergencies (Toriumi et al. 2013).

Previous work focused on modelling the influence of *individual actors* in *single* media channels such as Twitter (Cha et al. 2010). Apart from a few studies such as the work of Argawal et al. 2012 the interdependencies between content in different online media channels is an underexplored area. To this end, we focus on the *relations* between individual *content items* in *different* types of online media and their role in information diffusion processes. Figure 1 shows an example of content spread via different online media types. We model information pathways that implicitly take into account time such that diffusion processes are not only characterized by the reached nodes but further by the speed

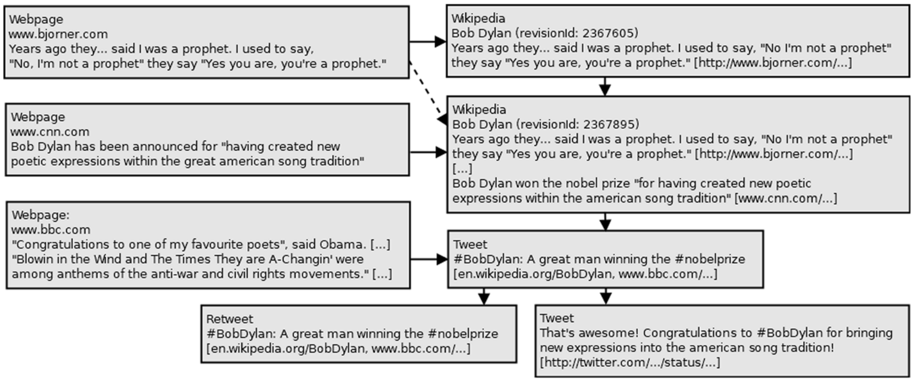

Webpage
www.bjorner.com
Years ago they... said I was a prophet. I used to say,
"No, I'm not a prophet" they say "Yes you are, you're a prophet."
Wikipedia
Bob Dylan (revisionId: 2367605)
Years ago they... said I was a prophet. I used to say, "No I'm not a prophet"
they say "Yes you are, you're a prophet." [http://www.bjorner.com/...]
Webpage
www.cnn.com
Bob Dylan has been announced for "having created new
poetic expressions within the great american song tradition"
Wikipedia
Bob Dylan (revisionId: 2367895)
Years ago they... said I was a prophet. I used to say, "No I'm not a prophet"
they say "Yes you are, you're a prophet." [http://www.bjorner.com/...]
[...]
Bob Dylan won the nobel prize "for having created new poetic
expressions within the american song tradition" [www.cnn.com/...]
Webpage:
www.bbc.com
"Congratulations to one of my favourite poets", said Obama. [...]
"Blowin in the Wind and The Times They are A-Changin' were
among anthems of the anti-war and civil rights movements." [...]
Tweet
#BobDylan: A great man winning the #nobelprize
[en.wikipedia.org/BobDylan, www.bbc.com/...]
Retweet
#BobDylan: A great man winning the #nobelprize
[en.wikipedia.org/BobDylan, www.bbc.com/...]
Tweet
That's awesome! Congratulations to #BobDylan for bringing
new expressions into the american song tradition!
[http://twitter.com/.../status/...]

Fig. 1 Example of interdependencies between contributions in different types of online media

of information uptakes, as described in previous works (Hecking et al. 2018; Hecking et al. 2018).

In this paper, we particularly investigate methods to identify groups of content contributions with a similar sphere of influence. This aims at discovering potential indirect influences between users. Moreover, it helps to identify groups of people with similar access to information while the groups may lack a direct relationship. Non-negative matrix factorisation is employed to reduce the possibly large and complex diffusion network to its basic underlying structure, facilitating an easier interpretation. Groups of contributions with similar positions in information cascades reveal who has similar information at similar points in time. A possible application is the identification of potential information biases by investigating which groups are reached or not reached by certain information pathways. Furthermore, this abstraction allows to observe diffusion processes on a group-level and to infer roles of users and contributions. Such roles can, for example, be *forerunner* posts that are taken up quickly by many others or *latecomers* which denote posts that are usually leaves of diffusion chains and take up information with high latency.

In summary, three subsequent tasks of the analysis of information diffusion using network analysis techniques are addressed. (1) Data collection from different online media channels and interrelationships between content items, (2) modelling of diffusion networks with special consideration of time, and (3) a method to summarise such complex diffusion networks into an interpretable structure by grouping nodes with similar positions in the network. The remainder of this paper proceeds as follows. “Background” section summarizes the background of this work and reviews related work. Our data harvesting approach is described in “Data collection” section and the analysis methodology in “Proposed approach for positional analysis” section. “Evaluation and applications” section demonstrates the utility of the proposed approach by applying it to real-world datasets. Finally, “Conclusion” section concludes the paper and highlights possible future research directions.

Background

Modelling information diffusion in networks

One central objective in the existing literature on diffusion and influence in complex networks is to find a subset of nodes that have the highest impact on the information reaching other nodes given a hypothesized diffusion model. This is known as the influence maximization problem (Kempe et al. 2013) and is especially important for applications such as viral online marketing.

Apart from such diffusion models, the empirical investigation of spreading processes is of particular interest in social media since advances in this direction contribute to a better understanding of the role of individual actors or social media platforms in the evolution of trends and opinions (Bakshy et al. 2012). Formal empirical analysis of information diffusion requires a clear conceptualization of the notion of an *information item*. Such *information items* may refer to a particular news story, rumor, or web-resource. Since discussions in social media can be ambiguous and diverse in language, it is sometimes not obvious whether two contributors refer to the same piece of information. Thereby, an important aspect is the level of granularity at which the conveyed *information* is analyzed. The notion of an information item can be operationalized on different levels of granularity, for example, based on topic models (Hong and Davison 2010), user-defined hashtags (Tsur and Rappoport 2012), or on a more fine-grained level based on n-grams or single terms (Aiello et al. 2013) (see Guille et al. (2013) for an extended discussion).

Other studies on information diffusion use the occurrences of concrete entities such as pictures or videos that are referred to and identified via URLs (Cao et al. 2015). The identification of information items on this basis is least ambiguous, but however, an exclusive concentration on URL sharing can be too narrow in many cases.

This work takes up ideas from meme tracking (Leskovec et al. 2009) as a method to identify multiple occurrences of different variations of short textual phrases across texts, which gives an adequate level of granularity by allowing variability of information items.

Apart from modelling and identifying information items, another challenge lies in inferring relationships between individual contributions. Adar and Adamic 2005 and later Gomez-Rodriguez et al. 2012 developed approaches to infer the most likely diffusion network given a sequence of ”infections“ of nodes. However, contributions in online (social) media often contain “citations” of other contributions, for example, hyperlinks to related information sources (Adar and Adamic 2005), or direct mentions of the original source as it is usual in Twitter (Taxidou and Fischer 2014; Cogan et al. 2012; Galuba et al. 2010). In this way, relationships become observable. In this work mixed-media information diffusion networks or ”citation networks“ are built from the combination of those observable relationships between content items.

Positional analysis of diffusion networks

Mapping nodes of a complex network $G_1(N_1, E_1)$ to nodes of a smaller network $G_2(N_2, E_2)$ with $|N_2| < |N_1|$, such that the edges E_2 reflect relational patterns between classes of nodes that occupy similar positions in G_1, is commonly referred to as blockmodelling (Doreian et al. 2004). The reduced network G_2 is a structural abstraction of G_1 and serves as an interpretable macro-structure, representing the relations of G_1 on a higher level. Blockmodels are often used to infer roles of individual nodes under the assumption that the individual behaviour of actors in social networks and abilities to act co-evolve with the network structure. It is important to distinguish blockmodelling from community detection (Fortunato 2010) which also aims at clustering nodes. However, while community detection aims at finding densely connected groups of nodes that are well separated from other groups, blockmodelling explicitly concentrates on modelling inter-group relations without requiring any inner-group links. In classical blockmodelling approaches, nodes are clustered based on their immediate neighbourhood, for example, by finding a homomorphism from G_1 to G_2 based on regular or structural equivalence (White and Reitz 1983).

In the case of information diffusion, positions of nodes in the network can be interpreted as roles in the diffusion process. Typical examples are citation networks of scientific publications. A special property of these networks is that they have an inherent notion of time. Since a publication can only cite a publication that has been published before, edges induce a partial order of the nodes, and consequently, the directed citation network cannot contain cycles. As mentioned before, the networks in this work have some commonality with citation networks. Since a contribution can only link back to already existing contributions they are also directed and acyclic as citation networks. One type of positional analysis in citation networks is the main path analysis technique (Hummon and Dereian 1989) that was originally developed to identify the main flow of information in citation networks. Here, the direction of edges is usually conceived as the direction of information flow, i.e. from the cited paper to the citing one. The main path then comprises of the edges that are most traversed by taking all possible paths from the source nodes (i.e. nodes with no ingoing edges) and to the sink nodes (i.e. nodes with no outgoing edges). This technique has also been adapted to interlinked revisions of different wiki articles (Halatchliyski et al. 2014) and social media networks (Hecking et al. 2018).

Another technique that has some commonality with our approach is directed acyclic decomposition of directed networks into components such that within a component all edges are reflexive and between two components edges are only allowed that point from the first component to nodes of the second [8]. The network between the components has to be acyclic. Originally developed to uncover hierarchical structures in networks, directed acyclic decomposition uncover clusters of actors whose access to information depends on other clusters in the sense that information is transferred between different components of a network.

The work presented in the following sections combines the idea of classical blockmodelling with modelling information diffusion networks as directed acyclic graphs. While in many cases blockmodelling methods can only be applied to dynamic networks if they are sampled into consecutive time slices (Hecking et al. 2017), the advantage here is that time is implicitly encoded in the network structure and no time slicing is necessary to incorporate temporal aspects in the role models. The work presented in this paper was inspired by the approach of Yu et al. (2005) who employed a special kind of non-negative matrix factorisation (Lee and Seung 1999) to identify a hierarchy of cluster affiliations of nodes in undirected networks which will be described in detail in "Decomposition of diffusion networks" section.

Data collection

The implemented data collection procedure starts with an initial search query based on hashtags, usernames, and keywords that is issued to the Twitter streaming API[1]. This initial query is dynamically expanded every 15 minutes based on the most frequently retrieved hashtags, usernames, and keywords within the last 15 minutes. As a second data source, Wikipedia's recent changes stream[2] is regularly checked for article updates that match the search criteria. Hyperlinks found in the retrieved data are recursively followed up to a depth of five and the items that match the search criteria are added to the set of items. The extracted data is stored in the DiscourseDB[3] format which was developed to store interlinked discourse items from various social media platforms in a unifying data model.

In the next step, networks are build in which nodes are the retrieved content items and the edges are denoted by URL references, if one item is a revision of the other (e.g. for Wikipedia articles), or if one is a retweet of another tweet. It is important to note that the edge direction is modelled to indicate the direction of information flow (from the referenced to the referencing contribution).

Since the initially assembled networks of contribution can be related to different subtopics of a news story, and thus, do not necessarily reflect the dissemination of a particular piece of information, a topic specific subgraph is extracted by querying all content nodes from DiscourseDB that contain a specified information item. As suggested by Leskovec, Backstrom, and Kleinberg (Leskovec et al. 2009), an information item is identified by similar phrases occurring in different contributions. In addition to contributions directly containing the phrase, contributions who refer to other contributions that contain the specified phrase (typically tweets functioning as vehicles to disseminate news articles) are collected as well. Two datasets were collected: The *Bob Dylan* dataset is based on web content containing phrases such as "Knocking on heaven's door", "Knocking on Nobel's door", etc., that were quite salient in social media at the time of data collection. The second dataset (*Schiaparelli*) relates to posts disseminating the information about the lost of the Schiaparelli lander during descent. The parameters for the initial search and the subgraph extraction used for the cased studies in this work are given in Table 1 were chosen based on experience and experimentation.

To avoid overestimation of the importance of contributions by the method outlined later, tweets that disseminate urls of articles almost directly after they were published, and thus were obviously generated by bots, were deleted. However, this happens only if they are not referenced, and thus, have no impact on the information diffusion process. If a contribution has no timestamp, which is often the case for web pages, the timestamp of the first reference to that contribution is taken as a proxy.

The analyses described in this paper were conducted on the largest weakly connected components of the resulting networks. Table 2 lists the composition of the two datasets.

Table 1 Timeframes, initial search parameters, and relevant phrases/items for subgraph extraction used in the case studies

Case study name	Timeframe	Initial search parameters	Relevant phrases/items
Bob Dylan	Oct. 12 – 17, 2016	nobel, #nobel, @Nobelprize, #NobelPrize, NobelPrize, nobel prize, #bobdylan, Bob Dylan, #bobdylan	Knockin on Heavens Door
Schiaparelli	Oct. 19 – 22, 2016	#EuropeanSpaceAgency, #ESA, European Space Agency, @esa, ESA, Exomars, #Exomars, Schiaparelli, #Schiaparelli, Mars, #Mars	#ESA, @esa, European Space Agency, #Schiaparelli, #ExoMars

Table 2 Number of nodes of the largest weakly connected components of the case study graphs after preprocessing

Case Study Name	Description	Number of Nodes			
		Twitter	Wikipedia	Youtube	Webpages
Bob Dylan	Bob Dylan wins Literature Nobel prize (Oct. 13)	169	1	2	13
Schiaparelli	ESA's Exomars mission, failed landing of the Schiaparelli probe on Mars (Oct. 19)	2230	2	0	11

Proposed approach for positional analysis

Time-weighted Katz coupling

The networks we are dealing with in this work are made of contributions (media content) on a topic linked by inter-references (e.g. hyperlinks). Chains of those references denote different information pathways. Furthermore, each contribution carries a timestamp which corresponds to the time when it was published. Thus, information pathways (or diffusion cascades) can also described in a temporal dimension with regard to the speed of diffusion. The position of a node (contribution) in an information diffusion network is thereby defined by the paths on which it can be reached in which time and the paths other nodes can be reached in a certain period in time. To model positions, modifications of the Katz centrality measure (Katz 1953) can be used to quantify the influence of contributions with respect to how many nodes can be reached and the time this takes (Hecking et al. 2018).

The approach relies on the notion of a weighted Katz matrix $\mathbf{K} \in \mathbb{R}^{|N|\times|N|}$, where N is the set of nodes (here content contributions) in a diffusion network. In the unweighted case, the elements $k_{i,j}$ of $\mathbf{K}$ give the number of directed paths between pairs of nodes $n_i, n_j \in N$, where longer paths are down-weighted by a factor $\alpha \in (0, 1)$. The Katz matrix can be calculated as given in Eq. 1.

$$\mathbf{K} = \sum_{l=1}^{\infty} \alpha^l \cdot \mathbf{A}^l = (\mathbf{I} - \alpha \cdot \mathbf{A})^{-1} - \mathbf{I} \tag{1}$$

Here, $\mathbf{I}$ is the identity matrix and $\mathbf{A}$ denotes the (weighted) adjacency matrix of the network. Consequently, the Katz matrix $\mathbf{K}$ gives the strength of the (indirect) influence between each pair of nodes. In directed acyclic graphs (DAGs), l can also be bounded by the diameter of the graph. In the general case, the Katz matrix can be computed using the inverse of the Laplacian of the adjacency matrix of a network, as shown on the right-hand side of Eq. 1. This is valid if α is smaller than the reciprocal of the largest eigenvalue of the adjacency matrix $\mathbf{A}$. Since in this work the Katz matrix is calculated only for DAGs, for which all eigenvalues are 0, any choice of α is possible. Experiments reported in Hecking et al. 2018 indicate that in the directed acyclic case α can be considered a scaling factor. The Katz matrix can also be calculated from weighed adjacency matrices taking into account the weight of edges on a path, which will be done in the following.

Since each contribution n_i carries a timestamp of its publication, time can be incorporated implicitly by setting the weight of an edge $(n_i, n_j) \in E$ in a diffusion network $G(N, E)$ to the inverse of their timestamp difference. The latency $\lambda(n_i, n_j)$ gives the time elapsed between the publication of n_i and a referencing contribution n_j. The inverse latency, and therefore the weight of an edge, is higher for edges that emerged due to quick take-up of information than for edges that link two contributions with a high temporal distance.

Since the values of edge latencies can span a wide range, and therefore, proper normalisation is required. Intuitively, small differences between small latencies should count more than small differences between very high latencies. For example, a difference of 1 second makes a large difference if the latency is 49 seconds. However, if the latency is 2 days the 1 second difference is less important. Therefore, we transform the raw latency (in minutes) λ into a normalized latency λ_{norm} using a sigmoid function, as given in Eq. 2.

$$\lambda_{norm}(n_i, n_j) = 2 \cdot \left(\frac{1}{1 + e^{\frac{-2\cdot\lambda(n_i,n_j)}{\mu(N)}}} - 0.5 \right) \tag{2}$$

This transformation ensures that any raw latency $\lambda(n_i, n_j)$ larger than the median raw latency $\mu(N)$ of the dataset is assigned a normalized latency $\lambda_{norm}(n_i, n_j)$ of 0.75 or higher. The $2\cdot(\cdots - 0.5)$ in the transformation is needed to map the function to the interval $[\,0; 1]$.

Decomposition of diffusion networks

Our approach described in the following is inspired by the work of Yu et al. 2005 for finding cohesive subcommunities in undirected graphs. However, the goal is not to optimize the separation between groups of nodes, but try to find interdependencies between groups in the sense of blockmodelling (see "Positional analysis of diffusion networks" section).

The Katz matrix $\mathbf{K}$ derived from the time-weighted adjacency matrix $\mathbf{A}$ of an information diffusion network $G(N, E)$ of web contributions described above, can itself be considered as the weighted adjacency matrix of a denser network $G_{katz}(N, E_{katz})$ that models direct and indirect influences between contributions. The weight of the edges in E_{katz} correspond to the time-weighted coupling of the incident nodes. More concretely, the more diffusion pathways from a node n_i to a node n_j exists and the lower the latency of the edges on this paths (indicating quicker information diffusion) the higher is the coupling between them.

The goal is to identify m classes of nodes $c_1, c_2, ..., c_m \in C$ that can be characterised in the sense that nodes having a high affiliation to the same classes (1) have similar access to information, i.e. they are reached by a similar set of nodes in a similar time span, and (2) have similar influence on succeeding contributions. Since each node in $G_{katz}(N, E)$ can have multiple roles in the diffusion process they are not uniquely assigned to classes but receive a weight for each class that indicates the strength of belonging.

Since conditions 1 (being reached) and 2 (reach others) are independent of each other, a node can have different affiliations to the node classes regarding in- and outgoing relations respectively. To address this consideration, each class c_i is modelled to have two sides c_i^- and c_i^+, where the negative side c_i^- refers to the incoming influence of nodes and the positive side c_i^+ refers to their outgoing influence. For example, two nodes can be reached by information items on completely different pathways (i.e. having no common predecessor) but when they further disseminate the information, they reach the same succeeding nodes in similar time. In this case they would belong to different classes c_i^- and c_j^- regarding their ingoing relations but they would have a common affiliation to a class c_k^+ with respect to their outgoing relations. This accounts for the duality of roles (or positions) nodes can have in diffusion networks, which has not been yet considered explicitly in related works on positional analysis mentioned in "Background" section.

Affiliations of nodes to the positive and negative sides of node classes can be modelled as a weighted and directed bipartite network, where N is the set of contributions in $G_{katz}(N,E)$ and C is the set of node classes (clusters). The weight of a directed edge from node n_i to class c_l, $w(n_i,c_l^+)$, corresponds to the strength of the affiliation of node n_i to c_l^+, and the weight of a directed edge from class c_l to node n_i $w\left(c_l^-,n_i\right)$ corresponds to the node's affiliation to c_l^-. An example can be seen on the right-hand side of Fig. 2. A node n_i for which $w\left(n_i,c_l^+\right)$ is high, should have a high influence on many nodes n_j for which $w(c_l^-,n_j)$ is high. More concretely, the class nodes in such bipartite network as given in Fig. 2 (right) summarise the relationships between nodes in the original diffusion network (left).

The adjacency matrix $\mathbf{A_B} \in \mathbb{R}^{(|N|+|C|)\times(|N|+|C|)}$ of this bipartite network has the typical form given in Eq. 3, where $\mathbf{W} \in \mathbb{R}^{|N|\times|C|}$ contains the weights $w\left(n_i,c_l^+\right)$ and $\mathbf{H} \in \mathbb{R}^{|C|\times|N|}$ the weights $w\left(c_l^-,n_j\right)$.

$$\mathbf{A_B} = \begin{pmatrix} \mathbf{0} & \mathbf{W} \\ \mathbf{H} & \mathbf{0} \end{pmatrix} \tag{3}$$

Using the matrices $\mathbf{W}$ and $\mathbf{H}$, the directed bipartite network B can be projected into two directed unipartite networks $G_N(N,E_N)$ and $G_C(C,E_C)$ with adjacency matrices $\mathbf{W} \times \mathbf{H}$ or $\mathbf{H} \times \mathbf{W}$ respectively. The first case is illustrated in Fig. 2, where $\mathbf{W}$ and $\mathbf{H}$ are projected to the adjacency matrix of an unipartite network $\mathbf{K}$.

The basic idea in this work is to revert this process. The unipartite network with time-weighted Katz matrix $\mathbf{K}$ as the adjacency matrix of $G_{katz}(N,E_{katz})$ is known while the goal is to find the factor matrices $\mathbf{W}$ and $\mathbf{H}$ that best summarise the relational patterns in $\mathbf{K}$ by assigning nodes to node classes. Similar to Yu et al. 2005, this can be modelled as a non-negative matrix factorisation (NMF) problem where $\mathbf{K}$ is approximated by the product of the factor matrices $\mathbf{W}$ and $\mathbf{H}$ as depicted in Fig. 2. Note that in many cases there is no unique solution for and the found solution for deriving $\mathbf{W}$ and $\mathbf{H}$ and the original network can only be approximated.

One way to calculate a well-fitting decomposition of the matrix $\mathbf{K}$ is to minimize the Frobenius norm (c.f. Lee and Seung (2011)) given in Eq. 4.

$$\min_{W,H} ||\mathbf{K} - \mathbf{W} \times \mathbf{H}||_F^2 \tag{4}$$

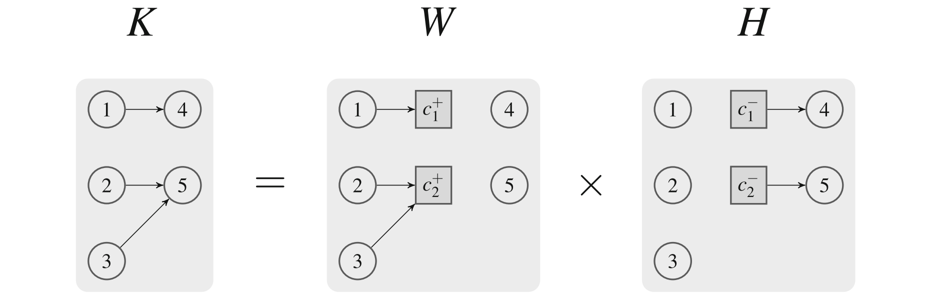


Fig. 2 The decomposition visualized for an example graph and $|C| = 2$

The term in Eq. 4 can be optimised by iteratively updating the factor matrices **W** and **H** according to the update rules given in Eqs. 5 and 6.

$$\mathbf{W} \leftarrow \mathbf{W} * \frac{\mathbf{K} \times \mathbf{H}^{\mathbf{T}} + \epsilon}{\mathbf{W} \times \mathbf{H} \times \mathbf{H}^{\mathbf{T}} + \epsilon} \tag{5}$$

$$\mathbf{H} \leftarrow \mathbf{H} * \frac{\mathbf{W}^{\mathbf{T}} \times \mathbf{K} + \epsilon}{\mathbf{W}^{\mathbf{T}} \times \mathbf{W} \times \mathbf{H} + \epsilon} \tag{6}$$

As shown by Lee et al. 2011, Eq. 4 is non-increasing under these update rules. It is important to note that $*$ denotes the element-wise multiplication of two matrices and the fraction denotes the element-wise division of two matrices. Since the networks in our case studies are directed and acyclic, there are always sources and sinks that result in zero rows or zero columns of **K** respectively. To avoid divisions by zero a small term ϵ is incorporated in the update rules.

There are different strategies for initialising the factor matrices **W** and **H** prior to the fist application of the update rules. In this work, the "NNDSVDa" initialisation strategy introduced by Boutsidis and Gallopoulos 2008 is applied, which is based on singular value decomposition (SVD) **K** with densification, as it yields good and stable results in our datasets after fewer iterations.

In this work, we used a slightly adapted version of this seeding strategy. All source nodes *Src* edges must have a fixed assignment to a class regarding their (non-existing) incoming relations and all sinks *Snk* should be associated to the same class regarding their (non-existing) outgoing relations. Therefore, two the negative respective positive sides of two different classes c_x^- and c_y^+ will be reserved for the sources and sinks (corresponding to the entries in row x of **H** and column y of **W**). More formally, this can be expressed as: $h_{i,src} = 1,\ if\ i = x,\ and\ 0\ otherwise,\ \forall src \in Src$. Respectively $w_{snk,j} = 1,\ if\ j = y,\ and\ 0\ otherwise,\ \forall snk \in Snk$.

Reduced diffusion network

The resulting approximation of the Katz matrix **K** by the factor matrices **W** and **H** do not impose a hard assignment of nodes to the positive and negative sides of classes. It rather assigns each node a weight of affiliation for each class. Furthermore, the nodes are clustered simultaneously with respect to their ingoing and outgoing relations, and thus, there can be overlaps in the positive and negative sides of different node classes that give an indication of information flow between different groups of nodes. For example, if many nodes with a strong affiliation to the negative side of a class c_i^- also have a strong affiliation to the positive side of another class c_j^+, this can be interpreted as these nodes receive their information from nodes in c_i^+ and provide information to nodes in c_j^-. This results in a strong coupling between c_i and c_j. This situation is depicted graphically in Fig. 3.

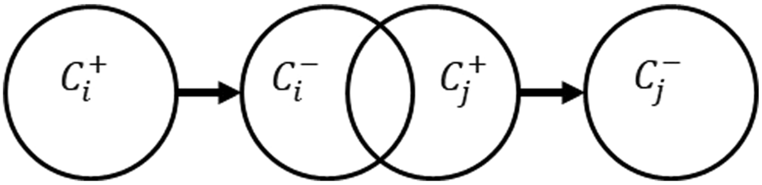


Fig. 3 Overlaps between the positive and negative sites of node classes

Based on the above considerations, a network of node classes $G_C(C, E_C)$ can be derived from the multiplication $\mathbf{D} = \mathbf{H} \times \mathbf{W}$, where $\mathbf{D}$ is its weighted adjacency matrix. The result can be considered as a blockmodel (see "Positional analysis of diffusion networks" section) of the original diffusion network in that it gives the relations between node classes with respect to information flow. In this regard, the strength of the relationship between classes c_i and c_j corresponds to the strength of the overlap between c_i^- and c_j^+. This graph is called the *reduced diffusion network* in the following.

Alternative approaches

In the literature, there are two related matrix factorisation methods that can also be applied for modelling dynamic and asymmetric relations between latent clusters. One of those is the RESCAL (Nickel et al. 2011) method, which can be considered as a relaxed version of the DEDICOM model (Bader et al. 2007). In RESCAL relational data is represented as a third-order tensor $\mathcal{X} \in \mathbb{R}^{m \times m \times l}$, where each slice $\mathbf{X_i} \in \mathbb{R}^{m \times m}$ of $\mathcal{X}$ represents relationships between the m objects. This can, for example, be a time slice of an evolving social network. In RESCAL objects can be categorised into k latent classes by finding a matrix $\mathbf{C} \in \mathbb{R}^{m \times k}$ denoting the associations of the m objects to k classes and matrices $\mathbf{R_i} \in \mathbb{R}^{k \times k}$ that indicate relationships between the classes in slice i, such that Eq. (7) is minimised:

$$min_{A,R} \sum_{i=1}^{T} ||\mathbf{X_i} - \mathbf{C} \times \mathbf{R_i} \times \mathbf{C^T}||_F \tag{7}$$

Since in this work time is captured implicitly in the edge weights of the diffusion network, RESCAL can be applied to the Katz matrix $\mathbf{K}$ instead of a tensor X and there is only a single relationship matrix $\mathbf{R}$ instead of multiple slices $\mathbf{R_i}$. The decomposition can be efficiently computed using algorithms based on alternating least squares that update the matrices A, R_i in alternating fashion by minimising the objective function in Eq. 7. In contrast to NMF, the matrices $\mathbf{C}$ and $\mathbf{R}$ are not necessarily non-negative. Thus, for better interpretability of the results a non-negative version of RESCAL was introduced by Krompass et al. 2013. This non-negative version is also used for comparison with the NMF based approach. For more details on the computation of the RESCAL factorisation we refer the reader to Nickel et al. (2011) and Krompaß et al. (2013).

Evaluation and applications

Estimating the number of node classes

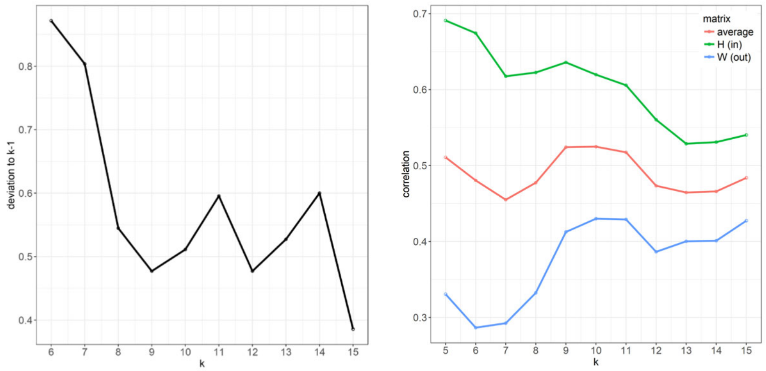
As in many clustering procedures, the choice of an adequate number of classes k is crucial for the utility of the outcome. In addition to using domain knowledge, more formal techniques can be applied. The quality, i.e. the error measured by Eq. 4, can not be used as a criterion since it tends to decrease with a growing number of classes. However, we can approximate how much new information can be added to the model by increasing k by testing to what extend pairs of nodes are still associated to the same clusters if k increases. This can me measured by projecting the node-classes affiliations given by the factor matrices $\mathbf{W}$ and $\mathbf{H}$ to associations between node pairs. For example, $\mathbf{M_{H_{k-1}}} = \mathbf{H_{k-1}} \times \mathbf{H_{k-1}}^T$ gives the associations between nodes based on the factor matrix of $k-1$ node classes. Node pairs with high affiliations to the same classes get higher values. In a similar way the associations for the other factor matrix $\mathbf{W}$ can be calculated.

Figure 4a depicts the average Frobenius deviation of the association matrices for successive k values taking the average of both association matrices $\mathbf{M_H(k-1)}$ and $\mathbf{M_W(k-1)}$ illustrated for the *Bob Dylan* dataset. It can be seen that more significant changes occur for $k < 8$, giving a lower bound for k.

As another heuristic more related to the actual task, one can estimate how well the clustering can be used to reduce the information diffusion network to an interpretable macro-structure that captures its underlying structure. As explained in "Reduced diffusion network" section, the weight derived for the connection between node classes c_i and c_j in the network of classes $G_C(C, E_C)$ (see "Decomposition of diffusion networks" section) indicates a high overlap of nodes that are most strongly affiliated with c_i^- (ingoing relations) and c_j^+ (outgoing relations to nodes in c_j^-). Consequently, for the network reduction to be meaningful this weight should go along with the average weight of the edges pointing from nodes most strongly affiliated to c_i^- to nodes having their highest affiliation for c_j^- in the original network $G_{katz}(N, E_{katz})$, which is denoted as $\bar{w}(c_i^-, c_j^-)$. In the same manner, the average strength of connections between nodes with the highest affiliation respectively to c_i^+ and c_j^+, $\bar{w}(c_i^+, c_j^+)$, should also correlate with the corresponding strength of the ties between classes in $G_C(C, E_C)$. Figure 4b depicts the Pearson correlation coefficients between the edge weights $w_C(c_i, c_j)$ in the network of classes $G_C(C, E_C)$, and, the corresponding $\bar{w}(c_i^-, c_j^-)$ and $\bar{w}(c_i^+, c_j^+)$ calculated from the original network. In addition to the information taken from Fig. 4a the number of node classes was set to 10 for the *Bob Dylan* dataset. For the *Schiaparelli* dataset the outlined heuristics suggest 6 different node classes.

Comparison with alternative approaches

Next, the reduced diffusion network derived by the approach described in "Reduced diffusion network" section ($\mathbf{D} = \mathbf{H} \times \mathbf{W}$) is compared with the non-negative version of the RESCAL decomposition described in "Comparison with alternative approaches" section. In contrast to the NMF based approach, RESCAL explicitly models the relationships between node classes in a relationship matrix $\mathbf{R}$ when decomposing the Katz matrix (see Eq. 7). In order to assess to what extend the strength of interdependence between node classes is captured by the matrices $\mathbf{D}$ using NMF decomposition or respectively $\mathbf{R}$ using



a Deviation to k - 1

b Cluster Correlations

Fig. 4 Characteristics of the NMF decomposition of **K** for different numbers of clusters *k*. **a** Deviation to k - 1. **b** Correlation between $w_C(c_i, c_j)$ and $\bar{w}(c_i^-, c_j^-)$

RESCAL, first each node is assigned to the class it has the strongest affiliation to according to **H** and respectively **W**, if NMF is used, or **C** (Eq. 7), if non-negative RESCAL is used. In a second step, for each pair of node classes c_i and c_j the average edge weight between nodes in c_i pointing to nodes in c_j is calculated (note that absent edges count with a weight of 0). Since a high weight for the pair c_i and c_j according to the models should go along with many high weighted edges between nodes in the corresponding classes a correlation analysis was performed to assess the quality of the models. The Spearman correlation coefficients ρ between the average edge weight between class pairs and the inter-class dependencies indicated by **D** (NMF) and **R** (RESCAL) are given in Table 3. For a fair comparison, both approaches were initialised with the same seed matrices. In the case of NMF nodes were assigned to clusters based on their outgoing relations indicated by the matrix **W** (NMF (c^+)) or their ingoing relations indicated by **H** (NMF (c^-)).

As it can be seen from Table 3, except in one case, the NMF approach captures density patterns between nodes having similar positions comparable or better than non-negative RESCAL. One reason for this observation can be that the explicit modelling of inter-class relationships applied in RESCAL introduces further complexity. Since two kinds of matrices (class affiliation **C** and class relationships **R**) are derived simultaneously in RESCAL, the results can be worse compared to the NMF model that captures inter-class relationships rather implicitly by allowing for overlaps between the positive and negative sides of classes c^+ and c^-. The observation that the strength of relationships between nodes classes is captured less good in the NMF model when nodes are partitioned regarding their c^+ affiliations can be explained by the order of the factor matrices in the matrice multiplication outlined in "Reduced diffusion network" section that maps negative to positive sides of classes and not vice versa.

Example I: Bob Dylan wins the Nobel prize

In the following the utility of the proposed approach is demonstrated along the *Bob Dylan* case study by decomposing the corresponding information diffusion network revealing different relational patterns. The network is depicted in Fig. 5. The upper colour of nodes represents their strongest associated node class regarding ingoing relations c_i^- and the bottom colour refers to the strongest associated node class regarding outgoing relations c_i^+. There are two source nodes, one is the Wikipedia article of *Bob Dylan* and another one denotes an article in a major news platform. By definition of the method, both are associated to c_1^- since they do not refer to any preceding contributions. However, they can be distinguished based on their outgoing relations. In part **a** of Fig. 5 it is shown that the Wikipedia article is associated to c_8^+ and is surrounded by Tweets news articles and one

Table 3 Correlations between the average edge weight between node classes and estimated node class interdependencies using different models

Dataset	Model	ρ
Bob Dylan	RESCAL	0.58
	NMF (c^+)	0.43
	NMF (c^-)	0.6
Schiaparelli	RESCAL	0.49
	NMF (c^+)	0.59
	NMF (c^-)	0.6

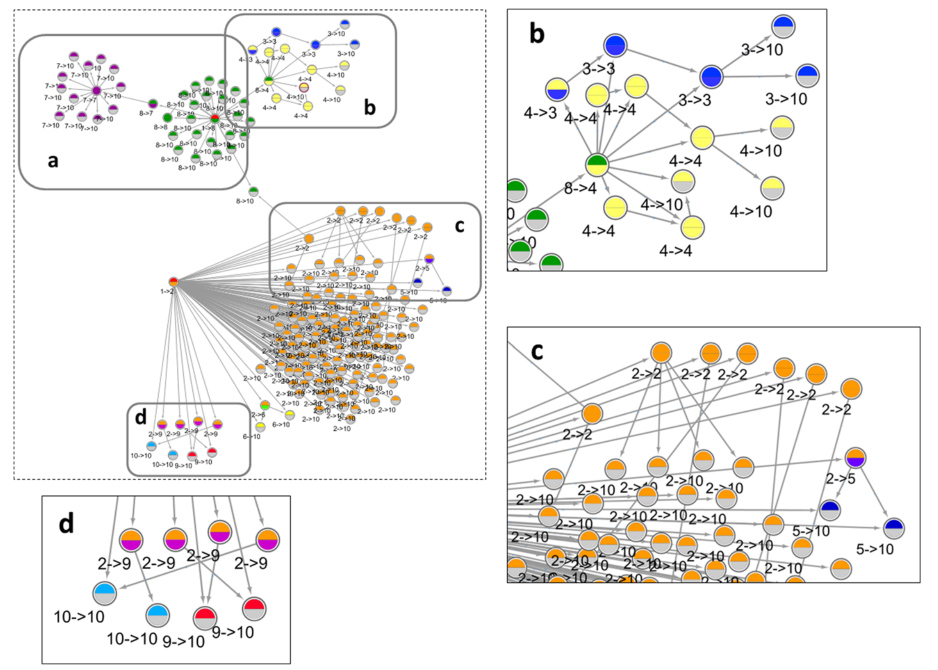


Fig. 5 Relational patterns in the Bob Dylan information diffusion network

YouTube video (c_8^-, c_{10}^+) that referred to it only a few hours after the update of the article (see 2nd row of Table 4). Since all these nodes are sinks they are assigned to c_{10}^+ by default. A more complex pattern can be seen in part **b** of Fig. 5. Here a single website (c_8^-, c_4^+) took up information from Wikipedia with a bit more delay but it was immediately taken up by tweets, for example in (c_4^-, c_4^+). This is possible since often news are simultaneously published in different platforms. The other variations in the node classes in part **b**, therefore, come from differences in the edge latency.

Parts **c** and **d** represent the diffusion cascades originating in the second source node (news page) associated to c_2^+ (6th row of Table 4). The depths of the information cascades in these parts is low since the news story was mainly propagated via tweets. Part **c** shows an example of 7 tweets that referred to the source page with only about an hour of delay on median (c_2^-, c_2^+). It is interesting to note that possibly because of the low latency they are not distinguished by the NMF decomposition with respect to their outgoing relations form the source node. They can be considered as influential in the sense that they were taken up by many other tweets in (c_2^-, c_{10}^+) however with higher delay and these tweets

Table 4 Characteristic pairings of node classes selected from the *Bob Dylan* dataset

c^-	c^+	Median time	Web	Tweet	Retweet	Wikipedia	Youtube
1 (*source*)	8	2016-10-13 13:03:05	0	0	0	1	0
8	10 (*sink*)	2016-10-13 22:00:42	11	10	0	0	1
8	4	2016-10-14 05:24:36	1	0	0	0	0
4	4	2016-10-14 05:24:36	0	5	0	0	0
4	3	2016-10-14 07:37:51	0	1	0	0	0
1 (*source*)	2	2016-10-13 19:00:42	1	0	0	0	0
2	2	2016-10-13 20:09:47	1	0	0	0	0
2	10	2016-10-14 11:41:11	0	105	6	0	0

were not further taken up by others (rows 7-8 in Table 4). In contrast to the quick spread of information from the source and partially mediated by tweets in part **c**, part **d** shows an example of a cascade with higher latency and less reach.

In general, it is interesting to see that the information originated in the Wikipedia page results in very quick and deeper diffusion cascades compared to the information pathways with a news page as source while the reach is considerably lower.

The node class information can additionally be used to map the network to a reduced diffusion network $G_C(C, E_C)$ as described in "Reduced diffusion network" section. The result is depicted in Fig. 6. The edge with corresponds to the strength of the coupling between node classes which results from the strength of the affiliations of nodes to the corresponding classes. For example, if many nodes with a strong affiliation to the negative side of a class c_i^- also have a strong affiliation to the positive side of another class c_j^+, this results in a strong coupling between c_i and c_j. The colours correspond to the colours used for ingoing node classes in Fig. 5. Class **1** again represents all nodes with no ingoing relations and class **10** the nodes with no outgoing relations. For this reason, class **10** can only be isolated in the mapped network since it is not possible that nodes receive links from class **10** nodes. It can be seen that the proposed methods produce a reduced diffusion network that reflects the different patterns and parts outlined about very well. At the top of the depiction the Wikipedia cascades with the deeper sub-cascades comprising of node classes **3** and **4** are mapped (part **b** in Fig. 5). The bottom nodes correspond to the flat cascades that mainly made by tweets that can be distinguished based on the latency of information takeups (parts **c** and **d** in Fig. 5).

Example II: Schiaparelli Mars lander lost during decent

The results for the *Schiaparelli* dataset show a different diffusion pattern. In contrast to the *Bob Dylan* case, the contribution types are mostly on Twitter. The news on the failure of the Schiaperelli landing first appeared on the web, it was quickly taken up by a large amount of tweets which results in several star-like structures with web pages in the center and tweets in the periphery. Since the original network is too large to visualise,

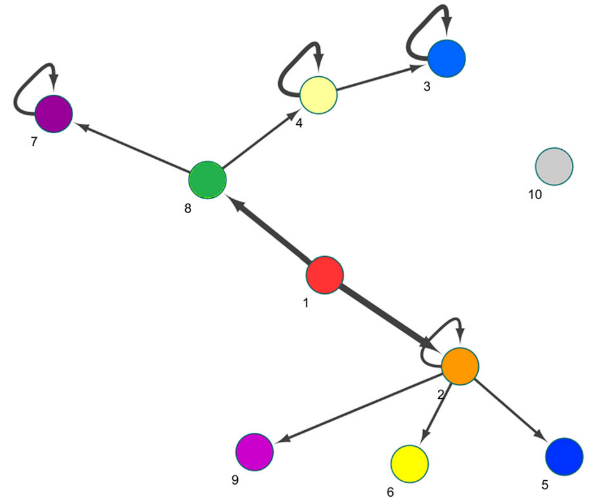


Fig. 6 The reduced diffusion network of node classes for the *Bob Dylan* case study

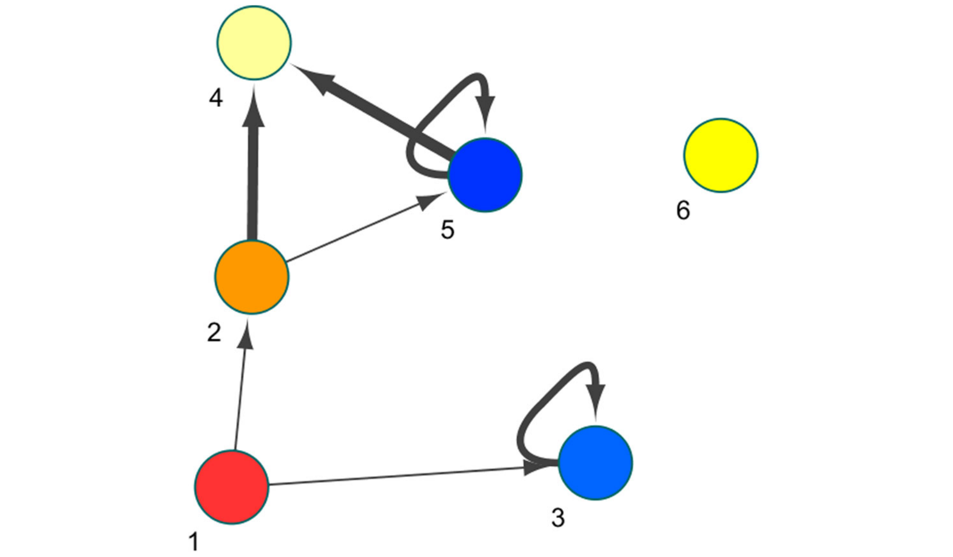


Fig. 7 The reduced diffusion network of node classes for the *Schiaparelli* case study

only the reduced diffusion network for the *Schiaparelli* case is depicted in Fig. 7, which reflects the structure of the original one. Webpages which constitute the main sources of information are assigned to the negative side of class c_1^- . Due to retweeting some small sub-cascades witing these stars emerge. The vast majority of contributions which are tweets in Twitter are mostly sinks and can be distinguished by their associations to the negative sides of node classes that reflect the magnitude of latency of information take ups from the sources.

The information regarding the node classes in the *Schiaparelli* dataset are subsumed in Table 5. It is interesting to see that the much of the information flow between the sources in class c_1^- and the sinks that can be found in c_3^-, c_4^- is only dependent on one or very few important tweets that can be found in c_3^-. This tweet was published by the Twitter account of a major news agency distributing a picture from their web page showing the Schiaparelli Mars lander after the lost during decent had been confirmed by the ESA. This tweet was taken up by other tweets very often, and thus, can be considered as a pivotal moment in the dissemination of the news story in social media. The indirection between c_2 and c_4 over c_5 results from a single tweet (6th row of Table 5) which was an immediate reaction to one of the tweets in in c_2^-. This might be possible through the use of specific apps that help users to disseminate content by copying and re-posting of content.

Table 5 Characteristic pairings of clusters selected from the *Schiaparelli* dataset

c^-	c^+	Median time	Web	Tweet	Retweet	Wikipedia	Youtube
1 (*source*)	2	2016-10-20 14:11:53	1	0	0	0	0
1 (*source*)	3	2016-10-19 21:33:25	3	1	0	0	0
2	4	2016-10-20 11:00:03	0	1	2	0	0
2	5	2016-10-20 11:00:03	0	32	328	0	1
3	6 (*sink*)	2016-10-20 03:21:44	0	208	7	1	0
5	4	2016-10-20 11:00:03	0	1	0	0	0
4	6 (*sink*)	2016-10-20 12:58:02	0	29	348	0	0

Conclusion

This paper presented a novel approach for positional analysis in information diffusion networks that can be adapted by subsequent works to contribute to a better understanding of how news, ideas or opinions spread across different online information channels. The important considerations underlying the idea are: (1) Information spreads across different channels: Content items can differ in their nature and play specific roles in diffusion processes, as it is the case, for example, with news pages and tweets. Considering the interrelationships between information items in different channels yields a more complete picture on how information disseminates on the web and the importance of particular contributions. (2) The positions of nodes are not only given by their immediate neighbours but rather by connecting paths between node pairs. Our approach further takes indirect couplings between contributions into account by introducing the time weighted Katz matrix (see "Time-weighted Katz coupling" section). The weight of the coupling between to nodes decreases with, both, their distance in the diffusion network (structural dimension) increasing latency (time difference between two interlinked contributions) on the connecting path (temporal dimension). (3) The temporal dimension has to be taken into account in addition to pure structural analysis. Influence is given by the number of nodes that can be influenced. However, more influential contributions affect following contributions in a shorter period in time. At this, a measure for edge latency was introduced in "Time-weighted Katz coupling" section that turns the time difference between a later contribution to a previous contribution it refers to into normalised edge weights such that time is implicitly encoded in the network.

By applying non-negative matrix factorisation to the weighted Katz matrix of a diffusion network, relational patterns between certain classes of contributions induced by similar position in the information diffusion network could be derived. Using this approach to reduce a network to an interpretable macro-structure makes the different roles of content items more explicit and highlights typical diffusion paths between media types. An important aspect of the appraoch is that the model accounts for the duality of roles (sender and receiver) by distinguishing between incoming and outgoing influence of nodes, which has not been much addressed before. This can especially be useful to identify potential information bottlenecks and uncover hidden influences between groups of users in social media, and thus, can contribute to the development of new support mechanisms for online information management.

The presented work also has some limitations that have to be addressed in future works. The datasets are dominated by contributions on Twitter which is a results from the data collection procedure which uses Twitter contributions containing hyperlinks as seed to start crawling related content items. Thus, this procedure could be further advanced by using more initial datasources, for example, newsfeeds. Since edges can only be established between items that could be stored in DiscourseDB, the ovserved network is likely to be only be a subnetwork of the actual diffusion network. In order to make more sophisticated statements about the global structure of diffusion of news items, however, much larger and more comprehensive datasets is one of the main challenges. The harvested datasets are nonetheless considered to be well suited to outline the utility and specific properties of the proposed method for positional analysis in diffusion networks, giving reasonable examples for relational patterns of interconnected information items of different types.

In addition of taking observable relationships such as hyperlinks and retweets as the basis to create information diffusion networks, in future work links between contributions could also be inferred even if there are no explicit references. Following suggestions rooted in knowledge discovery approaches (Adar and Adamic 2005; Gomez-Rodriguez et al. 2012), hidden relationships, for example resulting from copying of content, could be detected and used to augment the network.

Endnotes

[1] https://dev.twitter.com/streaming/overview, as of 08/29/17

[2] https://www.mediawiki.org/wiki/API:Recent_changes_stream, as of 08/29/17

[3] http://discoursedb.github.io/, as of 07/01/18

Acknowledgements
We thank the anonymous reviewers for their valuable suggestions and comments on a previous version of this paper.

Funding
This work is supported by the German Research Foundation (DFG) under grant No. GRK 2167, Research Training Group "User-Centred Social Media".

Availability of data and materials
The datasets and software code used for the presented analyses are available in the GitHub repository https://github.com/hecking/positional-analysis-diff-net.

Authors' contributions
TH developed the general framework for positional analysis in diffusion networks with input from the other authors. TH and LS were responsible for data collection, implemented the algorithms, and performed the analysis. All authors wrote, read, and approved the manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Received: 5 April 2018 Accepted: 6 November 2018
Published online: 15 January 2019

References

Adar E, Adamic LA (2005) Tracking information epidemics in blogspace. In: Proc. of the 2005 IEEE/WIC/ACM Int. Conf. on Web Intelligence. IEEE Computer Society, Lyon, FR. pp 207–214

Agarwal N, Kumar S, Gao H, Zafarani R, Liu H (2012) Analyzing behavior of the influentials across social media. In: Behavior Computing. Springer, London. pp 3–19

Aiello LM, Petkos G, Martin C, Corney D, Papadopoulos S, Skraba R, Göker A, Kompatsiaris I, Jaimes A (2013) Sensing trending topics in twitter. IEEE Trans Multimedia 15(6):1268–1282

Bader BW, Harshman RA, Kolda TG (2007) Temporal analysis of semantic graphs using asalsan. In: 7th IEEE International Conference on Data Mining. ICDM 2007. pp 33–42. https://doi.org/doi:10.1109/ICDM.2007.54

Bakshy E., Rosenn I., Marlow C., Adamic L. (2012) The role of social networks in information diffusion. In: Proc. of the 21st Int. Conf. on World Wide Web. ACM, Lyon, FR. pp 519–528

Boutsidis C, Gallopoulos E (2008) Svd based initialization: A head start for nonnegative matrix factorization. Pattern Recogn 41(4):1350–1362

Cao C, Caverlee J, Lee K, Ge H, Chung J (2015) Organic or organized? exploring url sharing behavior. In: Proc. of the 24th ACM Int. Conf. on Information and Knowledge Management. pp 513–522

Cha M, Haddadi H, Benevenuto F, Gummadi PK (2010) Measuring user influence in twitter: The million follower fallacy. In: Proc. of the 4th International AAAI Conference on Weblogs and Social Media

Cogan P, Andrews M, Bradonjic M, Kennedy WS, Sala A, Tucci G (2012) Reconstruction and analysis of twitter conversation graphs. In: Proc. of the First ACM Int. Workshop on Hot Topics on Interdisciplinary Social Networks Research. ACM, Beijing, CN. pp 25–31

Doreian P, Batagelj V, Ferligoj A, Granovetter M (2004) Generalized Blockmodeling (Structural Analysis in the Social Sciences). Cambridge University Press, New York, NY, USA

Fortunato S (2010) Community detection in graphs. Phys Rep 486(3):75–174
Galuba W, Aberer K, Chakraborty D, Despotovic Z, Kellerer W (2010) Outtweeting the twitterers: Predicting information cascades in microblogs. WOSN 10:3–11
Gomez-Rodriguez M, Leskovec J, Krause A (2012) Inferring networks of diffusion and influence. ACM TKDD 5(4):21–12137
Guille A, Hacid H, Favre C, Zighed DA (2013) Information diffusion in online social networks: A survey. ACM SIGMOD Rec 42(2):17–28
Halatchliyski I, Hecking T, Goehnert T, Hoppe HU (2014) Analyzing the path of ideas and activity of contributors in an open learning community. J Learn Analytics JLA 1(2):72–93
Hecking T, Harrer A, Hoppe HU (2017) Discovery of Structural and Temporal Patterns in MOOC Discussion Forums(Kawash J, Agarwal N, Özyer T, eds.). Springer, Cham
Hecking T, Steinert L, Leßmann S, Masias V, Hoppe HU (2018) Identifying accelerators of information diffusion across social media channels. In: Network Intelligence Meets User Centered Social Media Networks. Lecture Notes in Social Networks. Springer, Cham
Hecking T, Steinert L, Masias VH, Ulrich Hoppe H (2018) Relational patterns in cross-media information diffusion networks. In: Proc. of the 6th Int. Conf. on Complex Networks & Their Applications. Springer, Cham. pp 1002–1014
Hong L, Davison BD (2010) Empirical study of topic modeling in twitter. In: Proc. of the First Workshop on Social Media Analytics. pp 80–88
Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of dna theory. Soc Networks 11(1):39–63. https://doi.org/doi:10.1016/0378-8733(89)90017-8
Katz L (1953) A new status index derived from sociometric analysis. Psychometrika 18(1):39–43
Kempe D, Kleinberg J, Tardos E (2013) Maximizing the spread of influence through a social network. In: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, Chicago, IL, USA. pp 137–146
Krompaß D, Nickel M, Jiang X, Tresp V (2013) Non-negative tensor factorization with rescal. In: ECML Workshop on Tensor Methods for Machine Learning
Lee DD, Seung HS (2011) Algorithms for non-negative matrix factorization. In: Advances in Neural Information Processing Systems. pp 556–562
Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. Nature 401(6755):788
Leskovec J, Backstrom L, Kleinberg J (2009) Meme-tracking and the dynamics of the news cycle. In: Proc. of the 15th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining. ACM, Paris, FR. pp 497–506
Nickel M, Tresp V, Kriegel H-P (2011) A three-way model for collective learning on multi-relational data. In: Proceedings of the 28th International Conference on Machine Learning. pp 809–816
Taxidou I, Fischer PM (2014) Online analysis of information diffusion in twitter. In: Proc. of the 23rd Int. Conf. on World Wide Web. ACM, Seoul, KO. pp 1313–1318
Toriumi F, Sakaki T, Shinoda K, Kazama K, Kurihara S, Noda I (2013) Information sharing on twitter during the 2011 catastrophic earthquake. In: Proc. of the 22nd Int. Conf. on World Wide Web. ACM, Rio de Janeiro, BR. pp 1025–1028
Tsur O, Rappoport A (2012) What's in a hashtag? content based prediction of the spread of ideas in microblogging communities. In: Proc. of the Fifth ACM International Conf. on Web Search and Data Mining. ACM, Dublin, IR. pp 643–652
White DR, Reitz KP (1983) Graph and semigroup homomorphisms on networks of relations. Soc Networks 5(2):193–234
Yu K, Yu S, Tresp V (2005) Soft clustering on graphs. In: Proceedings of the 18th International Conference on Neural Information Processing Systems. NIPS'05. MIT Press, Cambridge, MA, USA. pp 1553–1560