RESEARCH **Open Access**

Foundations of Temporal Text Networks

CrossMark

Davide Vega[*] (ID) and Matteo Magnani

*Correspondence:
davide.vega@it.uu.se
InfoLab, Department of Information
Technology, Uppsala University,
Uppsala, Sweden

Abstract

Three fundamental elements to understand human information networks are the individuals (actors) in the network, the information they exchange, that is often observable online as text content (emails, social media posts, etc.), and the time when these exchanges happen. An extremely large amount of research has addressed some of these aspects either in isolation or as combinations of two of them. There are also more and more works studying systems where all three elements are present, but typically using ad hoc models and algorithms that cannot be easily transfered to other contexts. To address this heterogeneity, in this article we present a simple, expressive and extensible model for temporal text networks, that we claim can be used as a common ground across different types of networks and analysis tasks, and we show how simple procedures to produce views of the model allow the direct application of analysis methods already developed in other domains, from traditional data mining to multilayer network mining.

Keywords: Network, Text, Time, Model, Temporal text network, Human information network

Introduction

A large amount of human-generated information is available online in the form of text exchanged between individuals at specific times. Examples include social network sites, online forums and emails. The public accessibility of several of these sources allows us to observe our society at various scales, from focused conversations among small groups of individuals to broad political discussions involving heterogeneous audiences from large geographical areas (Zhou et al. 2017; Nerghes et al. 2014).

This information is undoubtedly very valuable, as shown for example by the large revenues of big Internet companies and by its usage during political campaigns, but it is also very complex because of its joint textual, structural and temporal nature. To cope with this complexity, researchers have typically focused on either the topology of the network, as commonly done in Network Science, or the text exchanged among individuals, using methods from Computational Linguistics. In some cases time has also been taken into consideration as in, respectively, the fields of Temporal Networks and Temporal Information Retrieval.

However, despite this broad interest in human information networks, only a limited number of works have been developed to address text, network topology and time in an integrated way and using a common data model. In our opinion, this is partly a result of the over-specialization of today's academia, and the fragmented and discipline-specific

development of network research. Unfortunately, omitting any of the three basic elements of temporal text networks may lead to significant information loss and prevent a deeper understanding of the information system, as exemplified in the next section.

A motivating example

One typical usage of social media data in research is to study how information propagates online. In one of the many studies on this topic, the authors have analyzed different aspects of the propagation process considering the online reactions generated by the death of a well-known Italian TV anchorman (Magnani et al. 2010).

In Fig. 1 we have reproduced (a) the information propagation network, showing which posts contained information obtained by which others, (b) the text of some of the posts generated about this event, and (c) a temporal pattern indicating the number of comments per day.

While each of these pieces of information alone reveals something, putting them together into a temporal text network (Fig. 1d) we obtain a much more comprehensive understanding of the process. On the one hand, we can see that for the posts representing explicit attempts to propagate information (e.g., *Mike passed away*) publication time is
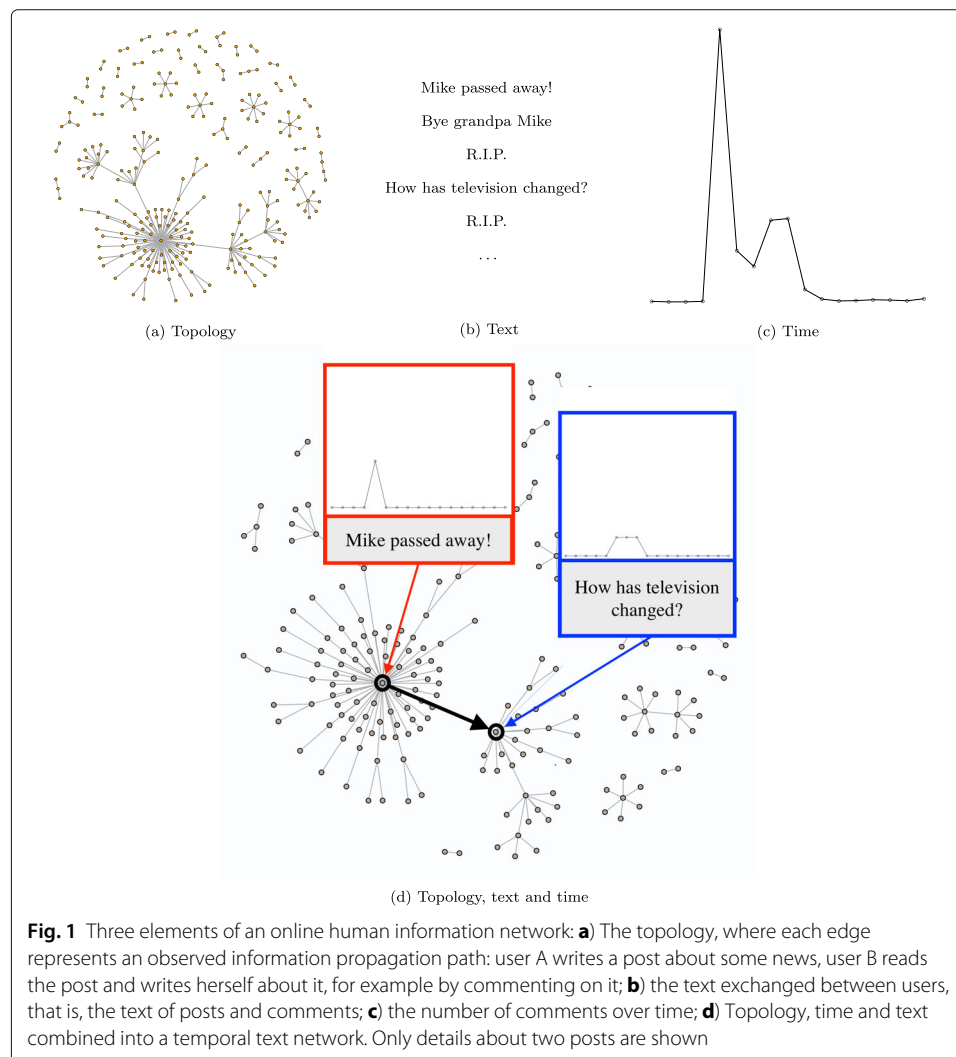
Fig. 1 Three elements of an online human information network: **a**) The topology, where each edge represents an observed information propagation path: user A writes a post about some news, user B reads the post and writes herself about it, for example by commenting on it; **b**) the text exchanged between users, that is, the text of posts and comments; **c**) the number of comments over time; **d**) Topology, time and text combined into a temporal text network. Only details about two posts are shown

fundamental to determine their success, and only the first of this type of posts generated a large and sudden burst of reactions in a very short time; on the other hand, conversational posts evolving from it (e.g., *How has television changed?*) can appear later and still create long but less dense chains of reactions. Other posts not present in the information propagation network neither explicitly give the news nor ask for an answer, generating no or few reactions, but still have the role of re-activating the information cascade so that even the latecomers can find a trace of it; some of these posts (e.g., *Bye granpa Mike!,* or *R.I.P.*) form what has been called an online mourning ritual.

In summary, time, text and topology together can lead to a deeper understanding of how this information network evolved into its current structure and how information propagated through it.

Contribution and outline

In this work we introduce a simple but expressive and easily extensible model for temporal text networks, and define two main approaches to analyze this type of data. We also show how existing primitive data manipulation operations for multilayer networks can be composed to easily construct new algorithms for temporal text networks.

Our claim is that such a model can play a similar role of other recent attempts to unify related areas of network science, such as multilayer networks, which have boosted research in already existing fields (e.g., multiplex network analysis) by showing that results in one area could be directly applied to other types of data now expressed using a uniform terminology and mathematical form. **Our objective is to define an essential model, with a minimal number of features, so that several existing models can be unified into it without a significant increase in model complexity**. We also believe that a unified model will promote the development of software libraries providing different data analysis functions for temporal text networks inside a single system, from centrality measures to community detection and generative models.

The article is organized as follows. In the next section we present an overview of related work, highlighting how a large amount of research has been produced to analyze human information networks. As the main objective of this article is to introduce a data model for temporal text networks, our overview of the state of the art focuses on the data models already introduced in the literature, to allow a precise comparison with our model. In "Modeling temporal text networks" section we define our model as a simple attributed bipartite network. We also show how this simple model can be used to represent many existing types of text-based interactions, such as direct messages, multicast and broadcast. In addition, we show how to express different types of information networks using our model, and how to extend it with additional features. Finally, we provide a detailed comparison of our model with the ones presented in the state of the art, showing how some existing models can be expressed using ours, while others can be obtained by applying some lossy processing to ours, e.g., replacing the exchanged text with a bag of words, a set of topics, a sentiment, etc. "Analyzing temporal text networks" section explains how the model can be used in data analysis. We show how the *direct* manipulation of the model can be complemented by two additional types of analysis: *continuous* and *discrete.* In the continuous case, time and text are treated as points in a metric space, and analysis operations are based on the computation of similarities between these points. In the discrete case, discretization operations (such as time slicing and topic modeling) are applied,

Vega *et al. Applied Network Science* (2018) 3:25

Page 4 of 26

encoding text and time into multiple discrete layers and enabling the direct application of the large number of methods already available for multilayer networks. In "A case study" section we present a practical example of our model and analysis strategies applied to Twitter data.

Related work

Our concept of temporal text network is a combination of text, network topology and time. In the literature there is a large number of models supporting one or more of these aspects, and the objective of this section is to characterize existing models from a common viewpoint. In this way, in the next section we will be able to provide a precise comparison between our proposal and existing work, showing that our model is more expressive but at the same time consistent with existing approaches, reusing existing modeling constructs when possible. In particular, we will show that we can express existing models using ours, but not vice-versa.

Notice that there are entire well-established disciplines developed to address text, network topology and time in isolation, and we do not review these here as they are widely covered by text books (Newman 2010; Baeza-Yates and Ribeiro-Neto 1999), described in numerous research papers (see for example (Blei et al. 2003) and subsequent extensions), and included in several software packages and systems. Instead we describe recent research efforts combining at least two of these aspects.

Table 1 presents a summary of the models selected for this review, also including our proposed model (core and extended temporal text network), organized according to four main criteria: (1) the type of graph used to represent the topological portion of the data, (2) the type(s) of nodes allowed in the graph, (3) the way in which text is represented in the model and (4) the way in which time is represented in the model. In "A comparison with the state of the art" section these criteria will be used for a comparison with our model. As our aim is to comprehensively list models, not papers, and the number of works using some of the models is very large, we have sometimes arbitrarily and unavoidably chosen a key set of references based on our knowledge and personal selection. Therefore, please notice that in the table we only indicate selected representative references; additional references are included in the text. Figure 2 complements Table 1 providing a visual intuition of the reviewed models and of the new models introduced in this article.
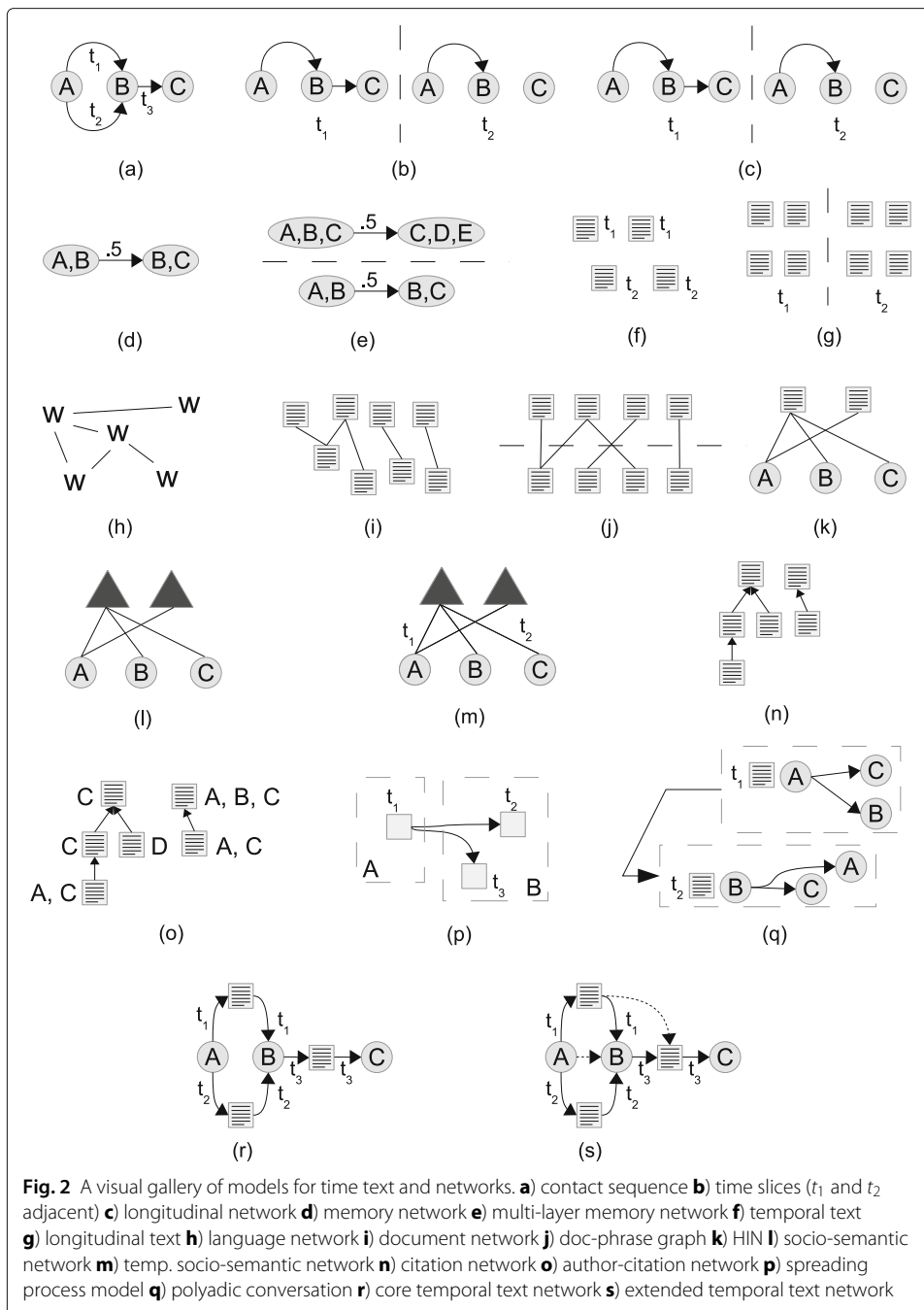
Time & topology

The most basic family of models including both time and topology is the *contact sequence* (Holme and Saramäki 2012; Gauvin et al. 2013). This is the most popular model for representing time and relations as a simple network structure. Mathematically, the model can be represented as a directed multi-graph $G = (V, E, T)$ with attributed edges. The set of vertices V represent actors (e.g., individuals, companies) and the set of edges E represent the interactions among the actors. When used in practice (Lambiotte et al. 2013; Gauvin et al. 2013; Paranjape et al. 2017) the duration of the interactions is sometimes considered negligible and hence represented as a single scalar $t \in T$, while in other occasions the temporal information is represented as time intervals $t = (t_s, t_e)$ indicating when the contact between two actors starts (t_s) and ends (t_e) (Viard et al. 2016). Contact sequences have been typically used to study information spreading (Lambiotte et al. 2013; Cheng

Vega *et al. Applied Network Science*   (2018) 3:25

Page 5 of 26

Table 1 Comparison of models representing two or more of the main attributes in temporal text networks

Name	Graph type	Node types	Text representation	Time representation	Refs.
Contact sequence	DMG	A	—	Mostly edges	(Holme and Saramäki 2012; Gauvin et al. 2013)
Time-slice	OML	A	—	Layers	(Mucha et al. 2010)
Longitudinal	OML	A	—	Layers	(Snijders 2005; 2014)
Memory	DG	A^n	—	Edges (implicit)	(Scholtes et al. 2014; Rosvall et al. 2014; Lambiotte et al. 2015; Peixoto and Rosvall 2017)
Memory (multilayer)	DML	$A^1 \cup \cdots \cup A^n$	—	Edges (implicit)	(Scholtes 2011)
Temporal text	—	X	Document	Vertices	(Brucato and Montesi 2014)
Longitudinal text	—	X	Document	Layers	(O'Connor et al. 2010; Dodds and Danforth 2010)
Language networks	G/DG	X	Word	—	(Sole et al. (2010))
Document networks	G/DG	X	Document	—	(Chang and Blei 2009; Menczer 2004)
Document-phrase graph	BG	$X \cup X$	Document, phrase	—	(Ren et al. 2017)
HIN	BG	$A \cup X$	BoW, concept, doc.	—	(Chang et al. 2009; Wang et al. 2017; Kralj et al. 2016)
Socio-semantic network	BG	$A \cup X$	Concept	—	(Roth 2017; Hellsten and Leydesdorff)
Temp. Socio-semantic network	BG	$A \cup X$	Concept	Edges	(Roth and Cointet 2010)
Citation network	DG	X	Document	Vertices	(Institute for Scientific Information et al. 1964; Batagelj)
Author-citation network	DG	$2^A \times X$	Document	Vertices	(White and Griffith 1981)
Spreading process	DG	$A \times X$	—	Edge (delay)	(Leskovec et al. 2007)
Polyadic conversations	DG	$A \times 2^A \times X$	Document	Vertices	(Magnani et al. 2012)
Core temporal text network	DBG	$A \cup X$	Document	Edges	
Ext. temporal text network	ML	$A \cup X$	Document	Edges	

The graph type is indicated as D: directed (undirected if D is not specified), O: ordered, G: Graph, MG: Multi graph, BG: Bipartite graph, ML: Multilayer graph. Node types indicate the domain of the nodes, and we distinguish between A (nodes used to represent actors) and X (nodes used to represent text-related objects). Given the variety of existing models, X is broadly used to represent full text documents, parts of it (phrases, words), other representations of documents such as bags of words (BoW), and also objects obtained by analyzing the text, such as concepts/topics. In this table we only indicate selected references

Vega *et al. Applied Network Science* (2018) 3:25

Page 6 of 26

Fig. 2 A visual gallery of models for time text and networks. **a**) contact sequence **b**) time slices (t_1 and t_2 adjacent) **c**) longitudinal network **d**) memory network **e**) multi-layer memory network **f**) temporal text **g**) longitudinal text **h**) language network **i**) document network **j**) doc-phrase graph **k**) HIN **l**) socio-semantic network **m**) temp. socio-semantic network **n**) citation network **o**) author-citation network **p**) spreading process model **q**) polyadic conversation **r**) core temporal text network **s**) extended temporal text network

et al. 2016), and existing concepts such as motifs and triadic closure have been re-defined to study the evolving structure of these networks (Paranjape et al. 2017; Viard et al. 2016; Kim and Diesner 2017).

Differently from contact sequences, where interactions are time-annotated one by one, other types of models use sequences of time-annotated graphs, where each graph is sometimes also called layer. In *time-sliced* models, also known as time-aggregated models, time is expressed as an interval and an edge indicates that an interaction has happened at some point during the time interval associated to the graph (Mucha et al. 2010). These models are typically obtained starting from a contact sequence and aggregating

edges by time. In *longitudinal networks* relationships about the same or similar actors are detected at different points in time (Snijders 2005; 2014). From a data modeling point of view, time slicing and longitudinal networks are very similar, and in practice the main difference lies in the nature of the time annotation associated to each slice, where in time slicing adjacent slices are typically associated with adjacent time intervals while in longitudinal network studies adjacent layers represent network snapshots obtained at specific points in time. Different types of time annotations are described for example in Batagelj and Praprotnik (2016).

Memory models provide a different view over a temporal network, where ordered tuples of two or more actors are represented as single nodes (Scholtes et al. 2014; Rosvall et al. 2014; Lambiotte et al. 2015; Peixoto and Rosvall 2017). For example, second order memory networks (Scholtes et al. 2014; Rosvall et al. 2014) can model the impact of one predecessor edge. For example, if an actor v_i is receiving one message from v_j and one from v_k, and is later sending a message to v_j and one to v_k, a contact sequence loses information on whether v_i is replying to v_j and v_k ($j \to i \to j$, $k \to i \to k$) or forwarding the messages ($j \to i \to k$, $k \to i \to j$). A first-order memory model will contain nodes for each pair of users and have an edge between two nodes if the corresponding pairs appear on consecutive paths. In our example, if v_i is replying we will have two edges in the memory model: $\left(\overrightarrow{ji}, \overrightarrow{ij} \right)$ and $\left(\overrightarrow{ki}, \overrightarrow{ik} \right)$, while if v_i is forwarding the messages we will have the edges $\left(\overrightarrow{ji}, \overrightarrow{ik} \right)$ and $\left(\overrightarrow{ki}, \overrightarrow{ij} \right)$. Higher order memory networks also exist (Lambiotte et al. 2015), although they are not as common, to represent causality effects between pathways consisting of 3 or more nodes. Deciding the order of the model is not trivial as specific patterns can be revealed only on a specific subset of memory models. To solve this problem, Scholtes (2011) introduced a multilayer memory network, composed of multiple memory networks of different order hierarchically connected between them (e.g., each node in the 2nd-order layer v_{ij} is connected with all nodes in the 3rd-order layer whose path $v_k lm$ contains the leg $\overrightarrow{ij}$, so $\overrightarrow{ij} \subseteq \overrightarrow{klm}$).

Time often plays an important role when networks are concerned, because networks often represent dynamical systems. However, in Table 1 we have only listed distinct data models explicitly providing time annotations. As an example, *growing network models* (Newman 2010) such as preferential attachment (Barabasi and Albert 1999) aim at explaining the observed topology of empirical networks based on how they evolve in time from an initial small network. Even if nodes and edges join the network one after the other, there is no explicit representation of time in the final model. Similarly, we have not listed papers about methods not explicitly introducing new data models, such as Lentz et al. (2013).

Time & text

Time is often present inside text, and commercial systems handling large human information networks from Google mail to common text messaging applications on smart phones can automatically identify the messages and annotate the text with temporal information.

In research, text and time are studied together in the field known as temporal information retrieval (Alonso et al. 2007; Kanhabua et al. 2015). This is an active area, also represented at the TREC conference where state-of-the-art information retrieval methods compete on various practical tasks. Time can be present in the text, as in the examples

Vega *et al. Applied Network Science* (2018) 3:25

Page 8 of 26

above or as metadata, expressed as absolute or relative time and it can also be specified in queries used to express information requirements (Brucato and Montesi 2014).

Another set of studies has focused on how text evolves in time, and in particular sentiment, with case studies ranging from tweets (O'Connor et al. 2010) to songs, blogs and presidential speeches (Dodds and Danforth 2010). Text and time are also studied across data sources, for example to correlate texts from online news to trends emerging in time series such as financial data (Lavrenko et al. 2000). However, no specific data model is used for this type of tasks, but only time-annotated documents (understood in a broad sense, including words, etc.) and time series.

Text & topology

Text and networks have been studied together in various areas, either without considering time or using networks to represent relationships between texts.

Models where nodes represent parts of a document have been used in structured information retrieval, which was a particularly active research area when hypertexts and markup languages became popular (Kotsakis 2002). Text is often contained inside some structure (e.g., a title, sections, sub-sections, etc.) and queries can be tuned to return specific parts of a document instead of a full one. As an example, if the searched keyword is contained inside Subsections 3.1 and 3.3 of a document, a query may return either the two subsections, or the whole Section 3, depending on the method.

More relevant for this article are document networks, that are graphs whose nodes represent text documents (Chang and Blei 2009; Menczer 2004). These network models can be classified in different groups depending on whether they include time or not; later in this section we refer to citation networks as a type of directed document network where time is also typically present. Text mining, and in particular clustering, can be applied to document networks to identify groups of documents that are similar not only because of their text but also because of their connections, as summarized in a recent article about clustering attributed graphs (Bothorel et al. 2015).

Several works have focused on networks extracted from text, and we can broadly classify them into models representing the text itself, aimed at characterizing language, and models representing actors and concepts mentioned in the text.

Networks where nodes represent words have been used to model both text documents and languages (Sole et al. 2010). For example, a document can be modeled as a network where words are connected by an edge when they are contiguous, or appear in the same sentence, paragraph, etc. Similarly whole languages can be modeled focusing on the relationships between words, as in WordNet or BabelNet.

With regard to the second class of models for networks extracted from text, Named Entity Recognition methods are typically used to identify the nodes and co-occurrence (or other language analysis approaches) to create edges among them (Diesner and Carley 2004; Chang et al. 2009). In this case, the output network connects different portions of a text document, or concepts extracted from the text.

A model that has been used to represent the relationships extracted from texts is known as heterogeneous information network (HIN) (Shi et al. 2017; Ren et al. 2016). HINs are defined as attributed directed graphs $G = (V, E, A, R)$ with an object type mapping function $V \rightarrow A$ and a link type mapping function $E \rightarrow R$, so that each object in the network (vertices and edges) belongs to a single type and if two edges belong to the same relation

Vega *et al. Applied Network Science*   (2018) 3:25

Page 9 of 26

type R, the two edges share the same starting object type as well as the ending object type. For example, HINs have been used in the past to model co-occurrence relations between entities (e.g., famous characters, sports, companies) in Wikipedia articles (Kralj et al. 2016). In Chang et al. (2009) vertices represent either famous characters from the text or bags of words, while the edges connect words that best explain the contexts where two or more famous characters appear together in the text. Document-phrase graphs as defined in Ren et al. (2017) are also HIN-based models, and more in detail probabilistic bipartite networks $B = (V, U, E, W)$ where the vertices in one partition V represent documents from a large document collection, the vertices in U represent salient phrases which are semantically relevant to one or more documents in V, and edges E indicate the relevance of each sentence for each document. HINs are not limited to represent relations within documents, text and concepts; but they can also model relations between actors and text. The most common use of HIN is actually to represent co-author or citation networks. In Wang et al. (2017), for example, the authors use an heterogeneous information network to describe the relations between scientific articles, their authors, and the venues where they were published.

One of the concerns recently raised against using methods from social network analysis to analyze social media is their intrinsic actor-centered approach (e.g., people, companies, stakeholders), focusing on social interactions without properly characterizing other aspects of the communication (Roth 2017). A similar argument can be used against the use of just Natural Language Processing or semantic networks (Sowa 2014).

Following this reasoning, a recent stream of research focused on combining structural and semantic data simultaneously, which led to the formalization of the socio-semantic network model (Roth and Cointet 2010; Roth 2017; Hellsten and Leydesdorff). Originally, socio-semantic networks were just bipartite graphs interconnecting *agents* (also known as actors in Social Network Analysis) with semantic objects called *concepts*, corresponding for example to terms, n-grams, or lexical tags.

During the last decade the socio-semantic network model has been extended to extract more valuable knowledge from social media. An illustrative example of such extension can be found in Hellsten and Leydesdorff () where the authors propose to combine the aforementioned social and socio-semantic networks into a single model. In short, they use a single matrix representation where the diagonal sub-matrices represent the relation between the same type of entities (agents and concepts) and the off-diagonal matrices represent the relation between different ones (agent/concept and concept/agent). From the point of view of data modeling, HINs are very related to socio-semantic network models, even though HINs have been introduced as more general modeling tools while socio-semantic networks have emerged and are used in a specific application context.

A final work worth mentioning in this class is Rosen-Zvi et al. (2004), where topic modeling is performed using an extended model considering not only the association between topics, words and documents, but also the association between documents and their authors. However, this has not been included in our summary table because it introduces a generative model to summarize the data in the form of parameters indicating the probability that a given actor produces a given set of words, but not to represent the empirical data showing which actors have written what text.

Vega *et al. Applied Network Science* (2018) 3:25

Page 10 of 26

Time & text & topology

Many works in the literature have dealt with time, text and topology using ad hoc models specifically designed to capture relevant aspects of specific platforms such as Twitter. For example, in Tamine et al. (2016) a communication network is built in three steps: (1) conversation trees are extracted from the dataset by inversely following the chain of Twitter user interactions (replies, mentions and retweets); (2) the trees are pruned based on the time elapsed between the root tweet and the overlap of tweets and participants in the tree; (3) finally, all trees are merged to generate a simple weighted graph of interactions between authors. A related model is the so-called polyadic conversation (Magnani et al. 2012), designed to describe user interactions in microblogging sites as a series of related conversations — also called polyadic interactions. A polyadic interaction is a tuple $i = (v, U, m, t)$ where $v \in V$ is the sender of the message $m \in M$, $U \subseteq V$ is the set of receivers and $t \in T$ is the timestamp of the communication act. A polyadic conversation is then defined as a chronologically ordered tree $G = (I, E)$ where I is a set of polyadic interactions and $E \subset I \times I$.

In Roth and Cointet (2010) a temporal model was used to compare the co-growth of two epistemic networks, a Twitter dataset and a set of related blogs, with the underlying social network of contacts. The temporal information attached to the edges of the network is, afterwards, used to compare the order of formation of epistemic and social communities.

Citation networks have received a lot of attention, and include text documents, directed edges between them and also time annotations (Institute for Scientific Information et al. 1964; Batagelj). In addition, when author co-citation analysis is performed (White and Griffith 1981), the underlying data model must also contain information about who authored which documents.

Information diffusion processes are often modeled including the diffused information item (meme, blog post, etc.), the actors propagating it, and the times of propagation. This is for example the case for the model used in Leskovec et al. (2007). However, the majority of these models do not use text to perform the data analysis, but (sometimes) to define the links between documents. Time can also be used to infer network structure based on the observation of propagation events. For example, the observation of a group of individuals repeatedly re-sharing common tweets in the same temporal order may suggest that these people are connected, and that information (tweets, in this case) passes through these hidden connections (Gomez Rodriguez et al. 2010). In Salehi et al. (2015) existing theoretical diffusion models for interconnected networks are reviewed, extending concepts in information diffusion to a multilayer model.

In order to preserve as much original information as possible, Šćepanović et al. (2017) use a more generic process to build the network, mixing techniques from social network and semantic analysis. In their work, the communication network is modeled as a simple, temporal graph using the Twitter "replies" to relate actors with each other. Then, they apply several semantic analysis procedures to generate supporting networks that describe the text-related features. A comparative analysis between the communication network and a subset of the semantic networks is used to study several aspects of the overall system such as semantic homophily and its evolution. However, from a modeling point of view text is not explicitly represented in this model, but coded inside the semantic layers. We will later use a related approach to exemplify how to use our model for data analysis.

Some attention has also been devoted to models describing co-evolutionary networks (Gross and Blasius 2008; Magnani and Rossi 2013). Some of these models allow the representation of a status associated to each node. Statuses can be used for example to represent the political affiliation of the person represented by the node. In growing network models, the status can influence the evolution of the network for example by increasing the probability that people will create connections with other individuals sharing the same political affiliation (Kimura et al. 2008; Lee et al. 2015). As for the case of simple network growing models, time is not typically kept at the end of the growing process, and in addition status has not been used to model text to the best of our knowledge. Therefore, we have not included these works in our summary table, even if we consider them potentially relevant for this field if extended in the future.

Modeling temporal text networks

In our opinion, a good model for temporal text networks should be general enough to be able to represent a wide range of systems, but also contain a minimal number of modeling constructs, to make the model easier to use and study. In other terms, a good compromise should be found between expressiveness and simplicity. In addition, given the large number of existing models that have been used for a long time to describe specific aspects of temporal text networks, we believe that both the modeling constructs and the terminology used in our model should be as aligned with previous work as possible. Following these design principles, we propose the following definition of temporal text networks:

Definition 1 [Temporal text network]
A temporal text network is a triple (G, x, t) where:

1. $G = (A, M, E)$ *is a directed bipartite graph, where, A is a set of actors, M is a set of messages, and $E \subseteq (A \times M) \cup (M \times A)$.*
2. $x : M \rightarrow X$, *where X is a set of sequences of characters (texts).*
3. $t : E \rightarrow T$, *where T is an ordered set of time annotations.*

and where the following constraints are satisfied:

1. $\forall m \in M, \text{in-degree}(m) = 1.$
2. $(a_i, m), (m, a_j) \in E \Rightarrow t(a_i, m) \leq t(m, a_j).$

In our model edge directionality indicates the flow of text in the network: $(a_i, m_j) \in E$ indicates that actor a_i has produced text m_j, while $(m_j, a_i) \in E$ indicates that actor a_i is the recipient of text m_j. Actors with out-degree larger than 0 are information producers, actors with in-degree greater than 0 are information consumers, and actors with both positive in- and out-degree are information prosumers.

Text is represented as a combination of a text container ($m \in M$), and a textual content ($x(m)$). As a consequence, actors in our model do not only generate text, but produce text messages. Two text messages (for example, two tweets, or two emails) may be different messages even if they contain the same text and have been exchanged between the same actors at the same timestamp.

The third key component of temporal text networks is the time attribute t. In our model, time is defined based on a generic set of ordered time annotations T. This enables the adoption of several ways of representing time: as an absolute date-time, as a relative

date-time, as a timestamp with an arbitrary format or as a discrete time interval if time has been sliced into time windows as it often happens when temporal networks are analyzed (See Table 1).
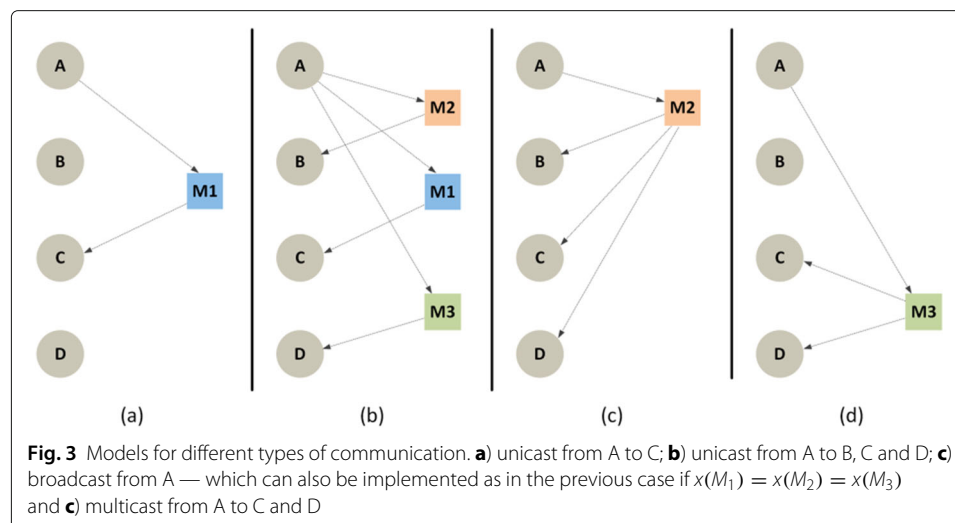
When writing about the model's elements, we will sometimes use a concise notation. For example, we will sometimes write an edge and its time together, as in: (a_i, m_j, t_q), where $t_q = t(a_i, m_j)$, and we will sometimes write a message by also indicating its sender, its recipients and its text, as in: $(a_s, m_j, \{a_{r_1}, \ldots, a_{r_n}\}, \text{"text"})$, where "text" $= x(m_j)$. Finally, when all the timestamps on the edges adjacent to a message are equal, we can also add a time to the previous notation, as in: $(a_s, m_j, \{a_{r_1}, \ldots, a_{r_n}\}, \text{"text"}, t_q)$.
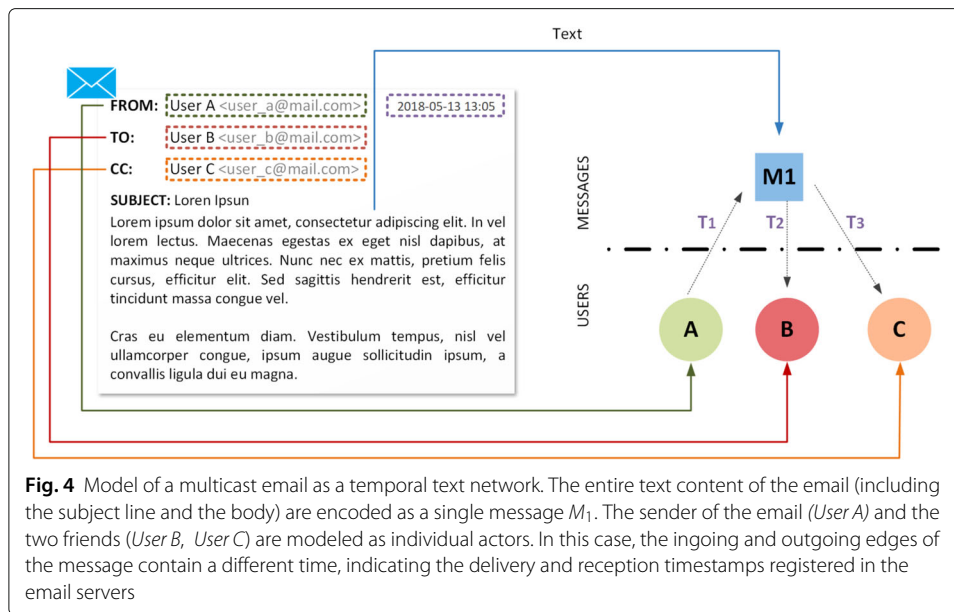
Applicability

While very simple, the model introduced above can be used to represent a range of different forms of communication and data from different sources. In particular, by explicitly dividing the network nodes into *actors* and *messages*, their relations implicitly carry more information. For example, whether the type of communication implemented by a message is unicast, multicast or broadcast is indicated by the out-degree of the message.

With **unicast** a message such as a handwritten letter is sent from a single source to a specific target. This form of written communication has been preserved to the present day through instant messaging services such as those offered by Twitter, Facebook Messenger or Whatsapp and, more traditionally, using the electronic email. Unicast communication allows to keep some text private between two actors, but it can have a large overhead if the same text must be sent to multiple sources because it requires an individual message for every recipient. In order to reach a larger population it is sometimes preferable to use **broadcasting** or **multicasting**. In the former, the message is transmitted to all possible receivers[1], while when the information is addressed to a group of people but not to all possible receivers, such as a post on a Facebook wall, the communication is called multicast. Figure 3 shows these different types of communication represented using our model.

Figure 4 shows an example of how a multicast communication through email can be modeled as part of a temporal text network. The resulting network includes the sender of the message (*User A*) and two other actors (*User B* and *User C*) who where explicit
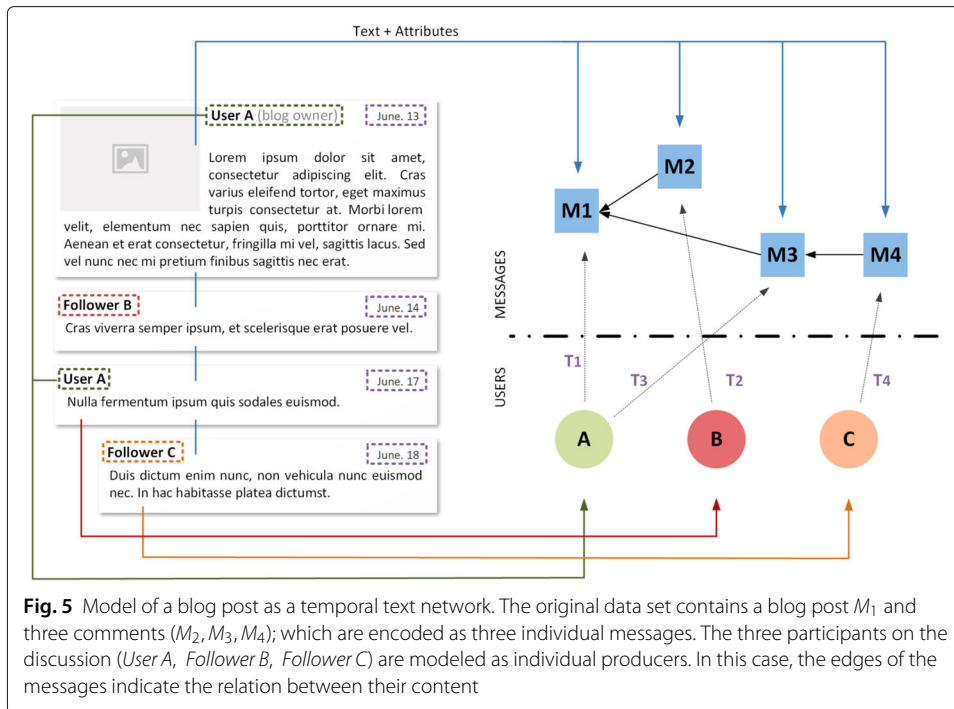

Fig. 3 Models for different types of communication. **a**) unicast from A to C; **b**) unicast from A to B, C and D; **c**) broadcast from A — which can also be implemented as in the previous case if $x(M_1) = x(M_2) = x(M_3)$ and **c**) multicast from A to C and D

Fig. 4 Model of a multicast email as a temporal text network. The entire text content of the email (including the subject line and the body) are encoded as a single message M_1. The sender of the email *(User A)* and the two friends *(User B, User C)* are modeled as individual actors. In this case, the ingoing and outgoing edges of the message contain a different time, indicating the delivery and reception timestamps registered in the email servers

recipients of the message. The fourth vertex $M_1 \in M$ represents the email and $x(M_1)$ corresponds to its text content (the subject line and the body content). In this case, the time attribute associated to each one of the edges represents the time when the message was delivered or received by the SMTP and POP3 servers allowing us not only to represent the communication flow, but also the effect of the channel and/or medium. Representing multiple emails as in the example above would lead to a full temporal text network.

In the next section we describe how to express other human information networks by extending our core model.
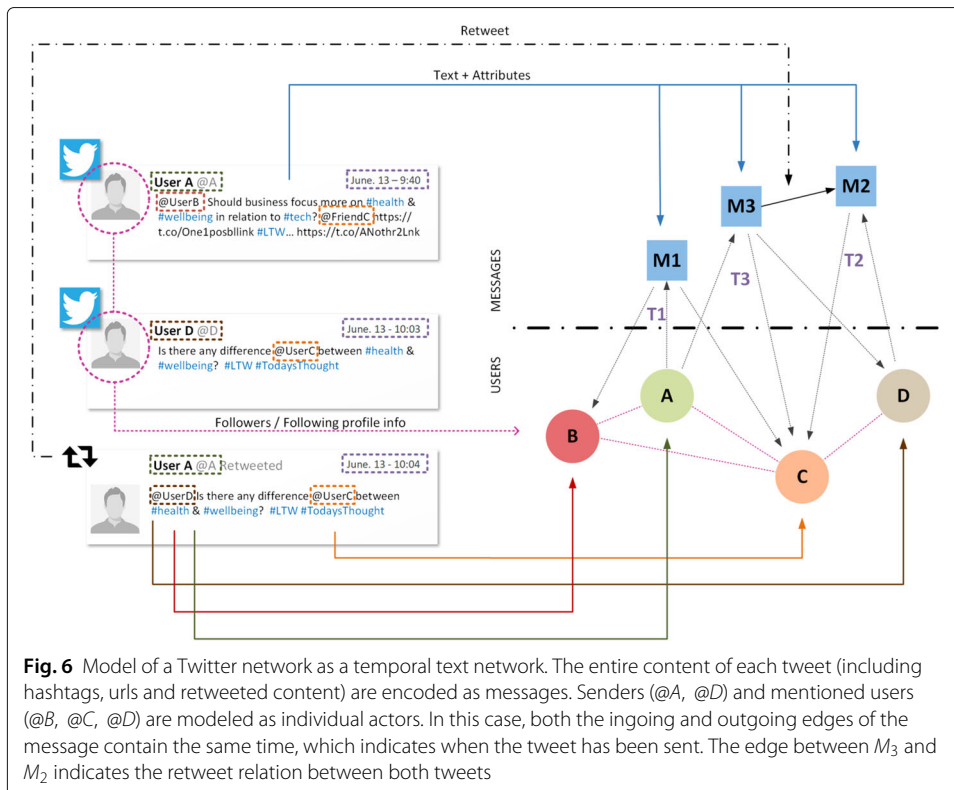
Model extensions

One of the design principles we used to define our model was simplicity, to make it tractable and general. On top of the basic model defined above, we can also easily add extensions to fit context-specific requirements.

With regard to the structure, we can straightforwardly add edges between messages to represent either information available from the data such as retweets on Twitter, or information deduced from the analysis of the data such as links indicating that one message is probably an answer to another, if we want to study information flows. Figure 5 shows, for example, the modeling process of a blog post M_1 and the associated comments from the readers $\{M_2, M_3, M_4\}$. In this particular case, we know the identity of each one of the authors, because they are authenticated in the web platform, but we do not know exactly who are the recipients of their comments. While we can assume by context that the blog post M_1 was read by follower B and that her message was then read by the blog owner A, it is uncertain what the third user (follower C) has read. We only know that the text produced by user C is a reply to the previous comment M_3, but we cannot infer if he has or has not read the previous messages M_1 and M_2. One possible way to model such scenario is to represent the relation between messages instead of the relation between messages *and* receivers. Similarly, in the example of Fig. 6 the edges between messages are used to represent retweets on a micro-blogging platform.

Fig. 5 Model of a blog post as a temporal text network. The original data set contains a blog post M_1 and three comments (M_2, M_3, M_4); which are encoded as three individual messages. The three participants on the discussion (*User A*, *Follower B*, *Follower C*) are modeled as individual producers. In this case, the edges of the messages indicate the relation between their content

As we discuss in the next sections, this type of extension would nicely fit our analysis framework where one main class of operations transforms the data into a multilayer representation. Similarly, we may add edges between actors indicating other types of relations relevant for the analysis of the human information network such as indirect recipients. Figure 6 shows the modeling process of Twitter as a temporal text network. Unlike the

Fig. 6 Model of a Twitter network as a temporal text network. The entire content of each tweet (including hashtags, urls and retweeted content) are encoded as messages. Senders (@A, @D) and mentioned users (@B, @C, @D) are modeled as individual actors. In this case, both the ingoing and outgoing edges of the message contain the same time, which indicates when the tweet has been sent. The edge between M_3 and M_2 indicates the retweet relation between both tweets

previous communication channels we discussed, in Twitter the recipients of the information are encoded in the text of the messages rather than being explicit in the metadata (e.g., the edge (M_1, B, t_1) exists because actor A mentions B in the first message of the data set). In addition, Twitter users can also see messages from other users they are following, which in our model is represented by the actor-to-actor relations. This difference between intra- and inter-layer relations allows us to differentiate between direct and indirect communication in many social platforms.

In our basic model x represents a generic string of characters over some alphabet, whose interpretation will depend on the source of the data and the context of the analysis. For example, while the symbol # usually denotes the start of a filtering tag in online social networks such as Twitter or Instagram, in other media sites it is just an acronym for the word "number". Therefore, for specific application contexts additional attributes can be added for example to messages by providing special information, such as the hashtags included in the text in the case of Twitter (See Fig. 6). In particular, we can think of having three types of information associated to each message:

1. The text, as in our basic model,
2. Metadata that is available in the specific data source used for the analysis, such as links to other resources (webpages, other tweets or multimedia content), like and retweet counts, or hashtags.
3. Additional information not directly available from the data source but obtained analyzing the text, for example through topic analysis.

Different types of temporal information have been used in existing works on temporal networks and temporal text analysis (See "Related work" section). For example, time can represent actions from the users such as the time when a message is posted and/or the time when it is read as we did in the Twitter example. Alternatively, times can be used to represent a physical property of the channel, as it happens in computer networks when there can be a transmission delay from the source to the destination of a message (See Fig. 4). Finally, time can also be associated to the message, indicating for example the time interval when the message exists. Furthermore, this information can be complete or incomplete, so that if only the initial time of the interval exists we must assume the message is still valid at the time of analysis as we did when we describe the blog posts; it can be private (accessible only to specific actors) or universally accessible by everyone.

A comparison with the state of the art

Our core and extended models of temporal text networks allow us to describe a variety of human information networks ranging from person-to-person email communication to complex interactions in social media sites. In "Related work" section we summarized other models from the literature, that have been used in the past to partially support similar scenarios. In this section we provide a comparative review between our models and the ones described in Table 1 and Fig. 2, emphasizing how they can be used to describe human information networks.

All models based only on time and topology (See Fig. 2a-e) do not include information about messages, documents or text. A simple extension adding a text attribute to the edges would still be less expressive than our model, because this simpler solution would not be able to differentiate between different types of communication such as unicast,

multicast and broadcast. These are instead allowed in our model exploiting the presence of nodes representing text messages, and thus justifying the adoption of a bipartite model instead of the simple graphs used in contact sequences. Single time annotations are also unable to distinguish between production/consumption or sending/receiving time. In summary, contact sequence models (Fig. 2a) can be expressed using our model by representing edges as edge-message-edge triples, but contact sequences cannot represent all the information that we can express using our model. Time-slices (Fig. 2b) and longitudinal models (Fig. 2c) can also be obtained starting from our model, as we do not make any assumption about how the time is represented on the edges. It is thus possible to represent both time-slices and longitudinal models as temporal text networks by just creating a new message m_j and a sequence of edges $(v_i, m_j, l), (m_j, v_k, l)$ for each original edge $e = (v_i, v_k, l)$ in the layer l of the sliced network. Finally, when only time and structure are concerned, memory models (Fig. 2d-e) are usually constructed from contact sequence models by aggregating the edges conditional on preceding pathways. While the original temporal information is partially preserved during the creation of the memory model, it is impossible to preserve more information from our temporal text network such as messages or network attributes. Therefore, we can think of our model as a way to represent raw and complete information about the temporal interactions and memory models as a way to emphasize information provenance. However, to represent provenance we need to allow edges between messages, and for this reason only our extended temporal text network model is able to express all the information present in memory models (in addition to text, multicasting and production/consumption times, as for all the other models not based on bipartite graphs).

The absence of relations makes it difficult to describe human information networks using just time and text (Fig. 2f-g), and despite their versatility to analyze text documents, strictly speaking none of the models only focusing on text and topology without actors (Fig. 2h-j) allows us to represent human-information networks as they do not contain any representation of the consumers and producers of the text. When also actors are represented, as in some HIN-based models (Fig. 2k) and in socio-semantic networks (Fig. 2l-m), our model adds directionality, which is necessary to represent text sender/receiver and producer/consumer relationships. Time is also not typically used in these models, but a temporal extension of existing HIN-based and basic socio-semantic models is straightforward and has in fact already appeared in the literature (Fig. 2m). The application of socio-semantic networks are also limited if compared with our model, as they contain already processed information (concepts) rather than text. With this we do not mean that our model is superior, as it can be useful to process the text into concepts, but this shows how we can go from our model to a socio-semantic model but not the other way round.

Citation networks and author-citation networks (Fig. 2n-o) can represent relationships between messages, and thus require our extended model to express their information. However, they cannot express communication, because even the more expressive author-citation network model (Fig. 2o) only focuses on the production of text. In particular, there are no edges between documents and authors, but only (implicit) edges between authors end documents. Spreading processes (Fig. 2p) also share the same limitations of either contact sequences or author-citation networks, depending on whether messages and/or authors are represented in the specific model, in addition of not (typically)

keeping the text content, which is however a minor problem as text can be easily added to the nodes representing the shared items. Compared with our core model, polyadic conversations (Fig. 2q) can express almost the same information: both can express unicast, multicast and broadcast relations between messages and actors, both differentiate between information producers and consumers and contain the raw textual information. However, while in our model each individual edge connecting messages and consumers can have a different temporal attribute, in the polyadic conversation model each polyadic interaction has one single temporal value.

Analyzing temporal text networks

One reason to adopt a common model instead of defining ad hoc models for each application is to reuse existing analysis methods. While our model can be analyzed *directly*, for example studying dynamical processes such as text propagation in a similar way as in our motivating example, we can consider other strategies. Here we define two more approaches that can be used to analyze temporal text networks: we call them *continuous* and *discrete*.

The practical benefit of using these two approaches is that instead of developing new algorithms the analyst can focus on defining mapping functions encoding the model in a way that fits the data and analysis at hand. Then, these functions automatically generate model views of which existing algorithms can be computed.
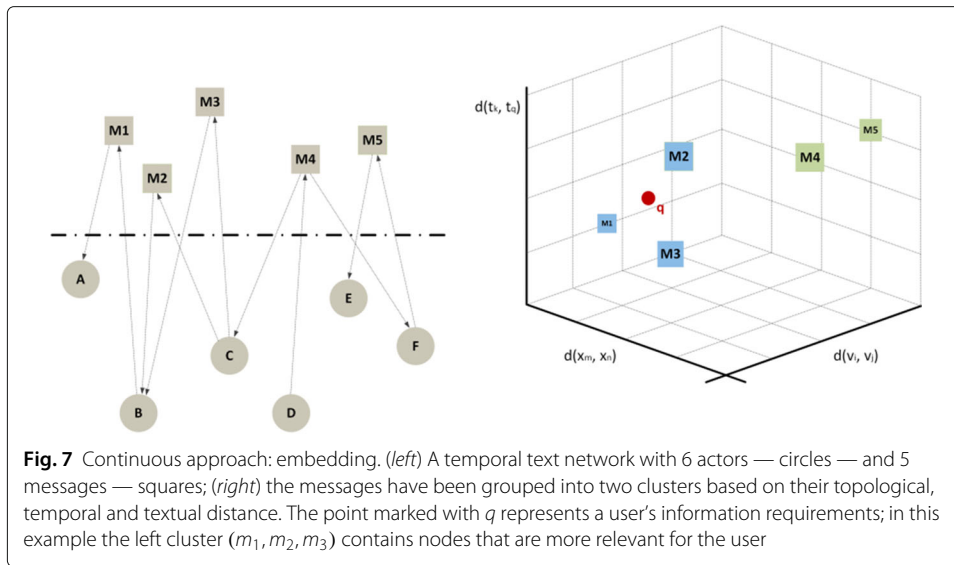
Continuous analysis

The main idea behind this approach is to map the elements of the network (e.g., actors, messages, content, etc.) into an asymmetric metric space. This means that it is possible to compute distances between them.

Once distances are available, one can directly reuse existing data analysis methods for metric spaces, such as traditional distance-based and density-based algorithms (k-means, db-scan, etc.). Distances can also be used to retrieve relevant information from large temporal text networks, specifying an information query as an element of the metric space and retrieving those elements that are the closest. We present an example of this last type of analysis in the next section.

The first way of doing this is to use a network embedding method (Goyal and Ferrara). While network embedding was initially defined for simple graphs, more recent algorithms can be directly applied to attributed graphs (Huang et al. 2017). Meanwhile, we foresee the definition of special versions of these algorithms that are specific for temporal text networks. Figure 7 shows an example of this first type of translation, where messages are the target of the analysis. The same approach can also be used to study other structures and elements in the temporal text network such as actors or combinations of actors and messages.

The second way to use the continuous approach is to directly define a distance function, without any explicit embedding into a coordinate system, so that the points form a metric space but have not an explicit position: only their relationships are defined. This approach is represented in Fig. 8.

The two approaches may look similar: in both cases algorithms use distances, which can be computed after an embedding or are directly defined in the distance matrix. In practice, however, there can be relevant differences. For example, after embedding it is
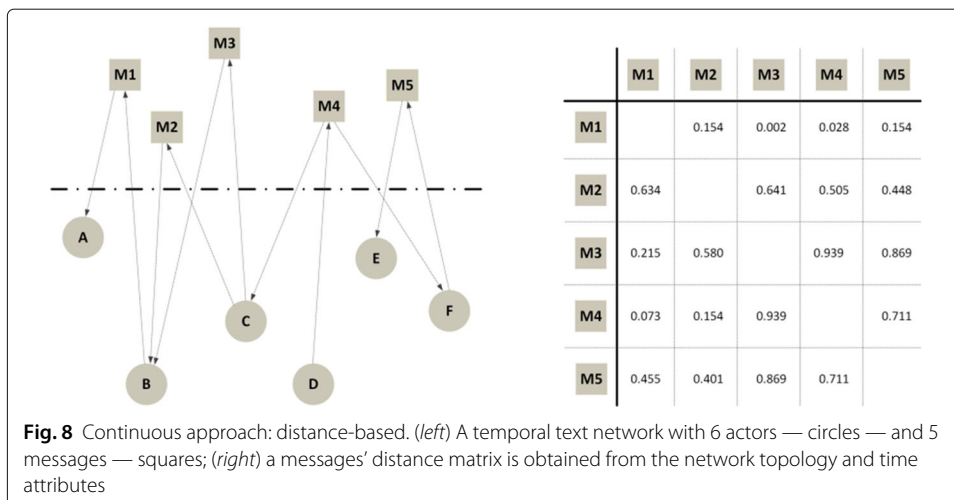
Fig. 7 Continuous approach: embedding. (*left*) A temporal text network with 6 actors — circles — and 5 messages — squares; (*right*) the messages have been grouped into two clusters based on their topological, temporal and textual distance. The point marked with q represents a user's information requirements; in this example the left cluster (m_1, m_2, m_3) contains nodes that are more relevant for the user

easier to index the data so that not all distances must be computed when algorithms are executed, leading to lower computation time. On the other hand, the direct usage of a distance function is more natural if distances are asymmetric, e.g., when $d(M1, M2) \neq d(M2, M1)$. Asymmetric distances often appear in temporal and directed networks, that are both features of our model.

Discrete analysis

The main idea behind this approach is to encode temporal and textual information into network structures, in particular layers in a multilayer network, so that methods from multilayer network analysis can be directly applied (Kivelä et al. 2014; Dickison et al. 2016). This can be done by defining a mapping function from time and text into a discrete set of classes that are relevant for the analysis. Then, topic-and-time-based user centrality, topic-and-time-based relevance, as well as community detection algorithms can be used. An example of this last type of analysis on real data follows in the next section.
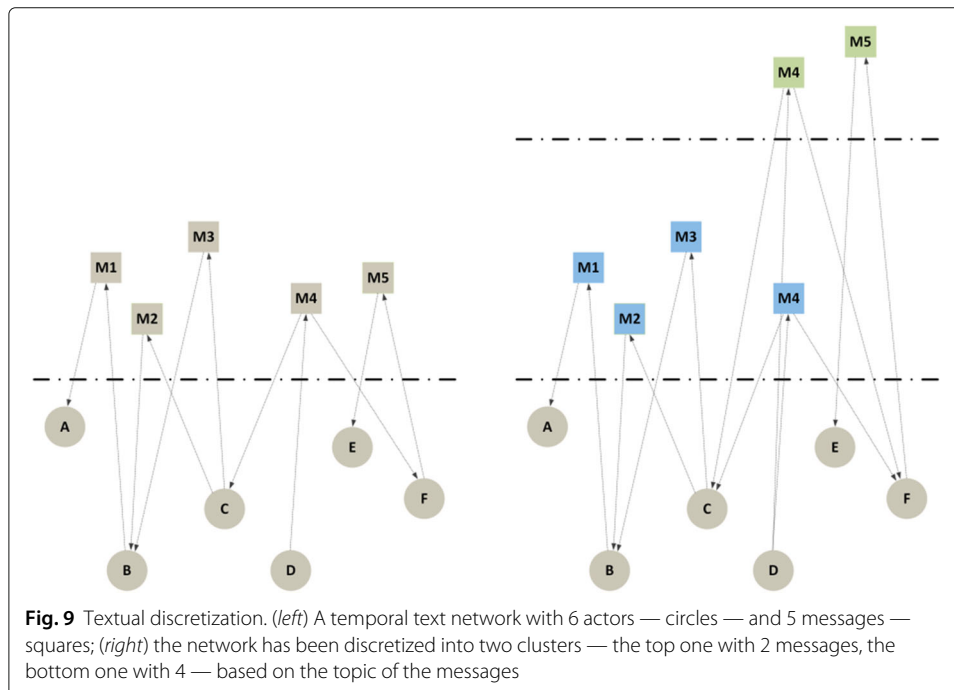
Textual discretization is typically performed using methods from Natural Language Processing such as topic, sentiment or semantic analysis. The main objective of the

	M1	M2	M3	M4	M5
M1		0.154	0.002	0.028	0.154
M2	0.634		0.641	0.505	0.448
M3	0.215	0.580		0.939	0.869
M4	0.073	0.154	0.939		0.711
M5	0.455	0.401	0.869	0.711	

Fig. 8 Continuous approach: distance-based. (*left*) A temporal text network with 6 actors — circles — and 5 messages — squares; (*right*) a messages' distance matrix is obtained from the network topology and time attributes
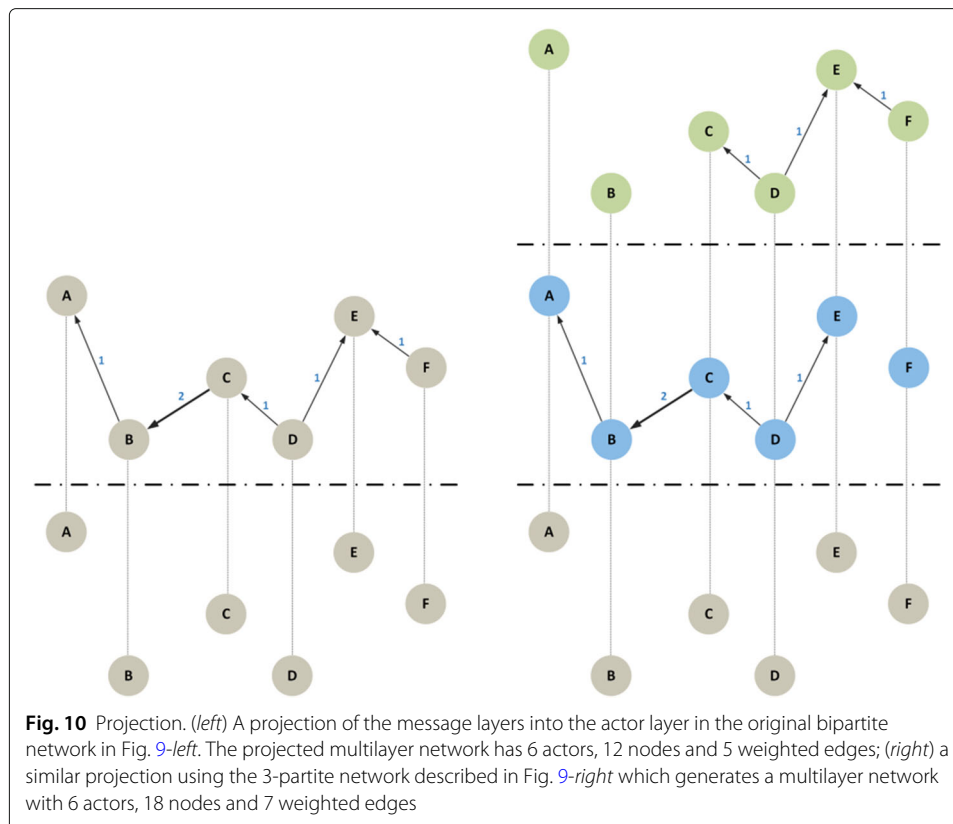
procedure is to group together messages whose contents have similar characteristics. *Time discretization* is apparently simpler, because only the cutting points between time slices must be indicated. However, also time discretization presents many options. First, there are often many ways of defining the cutting points, leading to different results. Second, after the cutting points have been defined there can still be different ways of distributing network structures into the slices. For example, if we want to discretize messages, we can place a message m_i in a specific interval (t_a, t_b) either if the incoming edge $e = (v_j, m_i, t)$ exists in the interval (t_a, t_b), if all the edges from/to m_i exist in the interval, if at least one of the out-going edges $e = (m_i, v_j, t)$ exist in the interval, etc. Finally, we use the term *multiple discretization* when both textual and time discretization are applied together to generate the different groups.

Under this procedure, our model would produce a k-partite network with one partition for each new cluster of messages and one partition for the actors. The procedure to generate such network is straightforward once the discretization function is defined. Figure 9 shows an example of textual discretization where the resulting 3-partite network contains the original layer of actors A, and two message layers with 2 and 4 messages each grouping together messages about the same topic. In this particular example, $x(M_4)$ was related to both topics, therefore the message M_4 appears in both layers. A similar network structure will emerge from time discretization.

An additional operation on multilayer networks that can be applied to the discretized data is projection, creating edges in one layer based on the information present in another layer. In the resulting multilayer network, a new edge $e_{ij}^{[l]} = (v_i, v_j)$ is created if there is a message m_k in the partition $l \in L$ of the original network with: a) an edge (v_i, m_k) from actor v_i to message m_k and b) an edge (m_k, v_j) from message m_k to actor v_j. Weights can also be added to the new edges, using various methods. Figure 10 shows one possible projection from the network in Fig. 9. In this example the content of the messages (and more in general also the time) are now encoded into the relations between actors.

Fig. 9 Textual discretization. (*left*) A temporal text network with 6 actors — circles — and 5 messages — squares; (*right*) the network has been discretized into two clusters — the top one with 2 messages, the bottom one with 4 — based on the topic of the messages

Fig. 10 Projection. (*left*) A projection of the message layers into the actor layer in the original bipartite network in Fig. 9-*left*. The projected multilayer network has 6 actors, 12 nodes and 5 weighted edges; (*right*) a similar projection using the 3-partite network described in Fig. 9-*right* which generates a multilayer network with 6 actors, 18 nodes and 7 weighted edges

The main advantage of using a projected multilayer network to analyze temporal text networks is the vast available literature that has targeted this type of data. In "Discrete analysis" section we use the approach described above together with a clustering algorithm for multilayer networks to find communities of actors discussing about the same topics during the same time spans.

A case study

In this section, we apply the model and approaches introduced in "Modeling temporal text networks" and "Analyzing temporal text networks" sections to a real temporal text network. In particular, we focus on using the discretization approach introduced in "Discrete analysis" section to analyze the formation and evolution of communities of actors and messages.

The objective of this section is two-fold. First, we want to give a concrete example of the abstract type of analysis described in the previous section. Second, we want to show in practice how a new type of analysis can be easily built as a composition of the transformations introduced in the previous section and an existing algorithm ("Discrete analysis" section).

Dataset

Our initial dataset consists of 247,399 public tweets with the hashtag *#iot* (Internet of Things) or some of its variants (e.g., #IoT, #IOT, etc.) automatically collected using the Twitter streaming API in June, 2017. The dataset contains mentions (tweets including *@username*), retweets (tweets starting with *RT @username*), other tweets that are neither

Vega *et al. Applied Network Science* (2018) 3:25

Page 21 of 26

mentions nor retweets, and the 51,369 users involved in the aforementioned communications. In order to improve the homogeneity of the collected data we further filtered our dataset by keeping only the tweets using at least one of thirty-two hashtags selected by domain experts as representative of main topics in this domain. This operation removed for example tweets containing the string *#iot* but not concerning the Internet of Things. In the following experiments we focus on the network obtained starting from the tweets containing mentions (about 5% of the initial tweets), built by coding each tweet as in Fig. 6.

The resulting temporal text network contains about one third of the users in the initial dataset (15,717) and the 13,210 messages exchanged between them (See Table 2). We call this the *original network*, and use it as a the starting point for both the following experiments.

Discrete analysis

Social interactions within a group of participants can form a community if they occur more frequently within the group than with other members of the network. In temporal text networks, those interactions are the result of the exchange of messages between actors. In this example we show how our model can be used to find communities of actors discussing about the same topics during the same weeks. Following the method described in "Discrete analysis" section we first transform our network to a multilayer network preserving information about interactions between users, topics and time, so that we can then apply an existing clustering algorithm.

The *discretized k-partite network* is built following the procedure explained in "Discrete analysis" section. In this particular example, we first split the original layer of messages using their hashtags as an indication of the topic, then we further discretize based on the week when messages are posted. The second discretization uses the posting time to create hashtag-week-specific layers.

Finally, we build the *multilayer network* by projecting each one of the layers containing messages into the actors' layer. Two actors in this network are connected in a given layer $L = (h, w)$ if at least one of them has sent a message to the other using the hashtag h during the week w. If multiple messages have been exchanged between two actors in the same layer, only a single edge is generated during the projection. At this step all edges are undirected and unweighted to fit the community detection algorithm we used. Table 2 describes the main properties of the original temporal text network, the projected k-partite and the final multilayer network used during the analysis.

Using the multilayer network and the clique percolation mechanism described in Mucha et al. (2010), we proceed to detect communities of actors across the whole network.
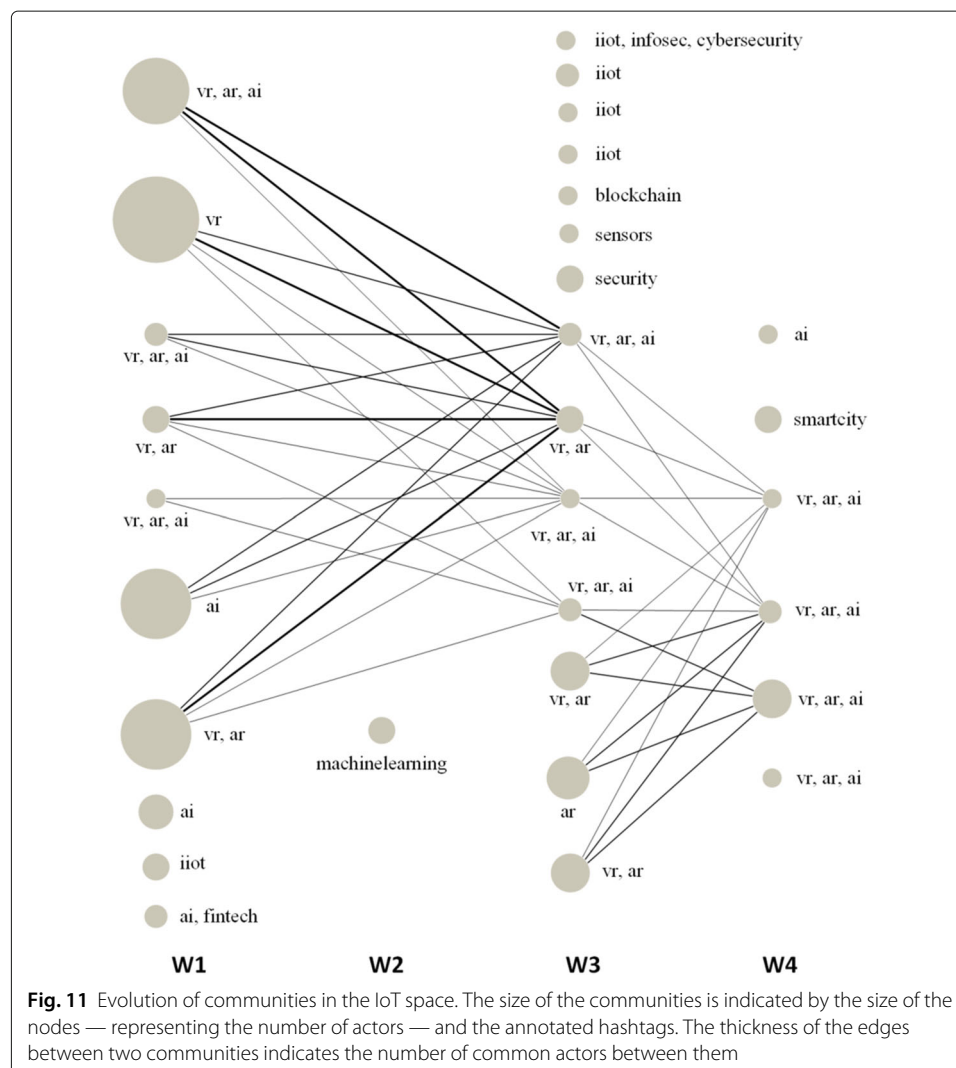
Figure 11 shows the communities with more than 3 actors formed in the multilayer network. Communities contain users and topics, and both users and topics can overlap across communities. The number of users is indicated by the size of the community, while

Table 2 Temporal text networks used in the case study and its basic properties: number of actors ($|A|$), number of messages ($|M|$), number of edges ($|E|$) and number of layers ($|L|$)

Networks	Type	$|A|$	$|M|$	$|E|$	$|L|$
Original	Bipartite	15,717	13,210	35,015	2
Discretized	k-partite	15,717	17,273	44,943	182
Projected	Multilayer	15,717	-	23,766	182

the layers representing the topics of interest of the actors are annotated next to each community. The smallest community in the diagram has 4 actors in the same layer, while the largest community contains 27 different actors and 3 layers. The edges between communities in different weeks indicate that at least one third of the users in the second community were also present in its predecessor. The thicker the line, the more users are shared between them.

We can observe that some of the hashtags, in particular *artificial intelligence* (#ai), *augmented reality* (#ar) and *virtual reality* (#ai), are very popular in the IoT space, with several groups of interest of different sizes forming around one or more of them. However, while the three topics are present across the whole month, the communities they form are very volatile. Only one of the smallest community with just 4 actors, for example, is preserved in time without changing its members or the topics they discuss. The largest communities formed during the first week, instead, disappear in week 2. Later on, some of the same users form new communities but with less members and a higher variance of topics. Less frequent hashtags such as #machinelearning, #security, #sensors, #smartcity and #blockchain also form groups of interest, usually smaller and with no or a few connections with the groups of users discussing the most common topics.

Fig. 11 Evolution of communities in the IoT space. The size of the communities is indicated by the size of the nodes — representing the number of actors — and the annotated hashtags. The thickness of the edges between two communities indicates the number of common actors between them

Overall these results suggest that the IoT space is very fragmented in this Twitter dataset. None of the found communities was big enough to become the main arena to develop a long-standing conversation on a specific topic. Instead, users organize themselves in smaller groups that change over time. Without combining topology, text and time we would find bigger communities, that would however include users talking about different things and at different times.

In summary, this example shows how a new analysis method can be easily constructed using our model and the approaches described in the previous section. In addition, also the results of this experiment highlight the value of using all the elements of the temporal text network in the analysis.

Discussion and conclusions

In this work we introduce a general model to represent temporal text networks based on the principles of expressiveness, simplicity and tractability. Our model is expressive and simple enough to encode the key components of human information networks (topology, time and text) into a single bipartite network, so that we can represent a range of different forms of communication and data sources spanning from postal services to online social media.

We additionally show how the model can be analyzed either directly or indirectly, to perform a variety of mining tasks. In particular, we define various transformations for two approaches that we call continuous and discrete. Using such transformations, we can map the data into existing models, allowing to reuse part of the machinery already developed to analyze complex data. While we do not describe each one of the possibilities enabled by our model in detail, in the experimental section we show two concrete experiments using the aforementioned transformations to analyze a set of communication messages exchanged in the Twitter platform during June 2017.

During the past century, the research community has demonstrated a huge interest in studying human information networks. As a consequence, researchers from different disciplines have devoted a considerable time to develop new models and methods to describe aspects of interest in this scenario. However, as we have shown in our review, there has been none or few successful attempts to unify the literature under a common framework: several models and algorithms have been proposed, but only for a subset of the aspects we consider in this article or they have been developed ad hoc to address a specific problem. So, results in one area cannot be directly applied to other types of data. We believe that our work can play a key role in the process of consolidating existing efforts from different disciplines under a common framework, in the establishment of a common terminology and in the development of new analytical software able to cope with the complexity of such data.

Endnote

[1] For simplicity we use the expression "all possible receivers" to refer to the community in which the information is spread, independent of whether the community is the whole Internet, the whole world or a set of members registered to a private site.

Vega *et al. Applied Network Science* (2018) 3:25

Page 24 of 26

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References
Alonso O, Gertz M, Baeza-Yates R (2007) On the value of temporal information in information retrieval. SIGIR Forum 41(2):35–41. https://doi.org/10.1145/1328964.1328968

Baeza-Yates RA, Ribeiro-Neto B (1999) Modern Information Retrieval. Addison-Wesley Longman Publishing Co., Inc., Boston

Barabasi A-L, Albert R (1999) Emergence of scaling in random networks. Science 286(5439):509–512. https://doi.org/10.1126/science.286.5439.509

Batagelj V Efficient algorithms for citation network analysis. CoRR cs.DL/0309023. http://arxiv.org/abs/cs.DL/0309023

Batagelj V, Praprotnik S (2016) An algebraic approach to temporal network analysis based on temporal quantities. Soc Netw Anal Min 6(1):1–28. https://doi.org/10.1007/s13278-016-0330-4

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993—1022

Bothorel C, Cruz JD, Magnani M, Micenkova B (2015) Clustering attributed graphs: models, measures and methods. Netw Sci 3(3):408–444

Brucato M, Montesi D (2014) Metric spaces for temporal information retrieval. In: European Conference on Information Retrieval, Vol. 8416 of LNCS. Springer Berlin Heidelberg. pp 385–397. https://doi.org/10.1007/978-3-319-06028-6_32

Chang J, Boyd-Graber J, Blei DM (2009) Connections between the lines: Augmenting social networks with text. In: Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '09. ACM, New York. pp 169–178. https://doi.org/10.1145/1557019.1557044

Chang J, Blei DM (2009) Relati onal topic models for document networks. In: Dyk DAV, Welling M (eds). AISTATS, Vol. 5 of JMLR Proceedings. JMLR.org. pp 81–88. http://dblp.uni-trier.de/db/journals/jmlr/jmlrp5.html#ChangB09. Accessed 21 May 2018

Cheng J, Adamic LA, Kleinberg JM, Leskovec J (2016) Do cascades recur? In: Proc. of the 25th international conference on world wide web. International WWW Conferences Steering Committee, Republic and Canton of Geneva. pp 671–681

Dickison ME, Magnani M, Rossi L (2016) Multilayer Social Networks. Cambridge University Press, Cambridge

Diesner J, Carley KM (2004) Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In: Causal Mapping for Research in Information Technology. IGI Global, Hershey. pp 81–108

Dodds PS, Danforth CM (2010) Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. J Happiness Stud 11(2):441–456. https://doi.org/10.1007/s10902-009-9150-9

Gauvin L, Panisson A, Cattuto C, Barrat A (2013) Activity clocks: spreading dynamics on temporal networks of human contact. Sci Rep 3:3099. https://doi.org/10.1038/srep03099

Gomez Rodriguez M, Leskovec J, Krause A (2010) Inferring networks of diffusion and influence. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '10. ACM, New York. pp 1019–1028. https://doi.org/10.1145/1835804.1835933

Goyal P, Ferrara E Graph embedding techniques, applications, and performance: A survey. CoRR abs/1705.02801. http://arxiv.org/abs/1705.02801. Submission Date: 12 Dec 2017

Gross T, Blasius B (2008) Adaptive coevolutionary networks: a review. J R Soc Interface 5:259—271

Hellsten I, Leydesdorff L Automated analysis of topic-actor networks on twitter: New approach to the analysis of socio-semantic networks. CoRR abs/1711.08387. http://arxiv.org/abs/1711.08387. Submission Date: 22 Nov 2017

Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97–125. https://doi.org/10.1016/j.physrep.2012.03.001

Huang X, Li J, Hu X (2017) Label informed attributed network embedding. In: Proc. of the 10th ACM International Conference on Web Search and Data Mining, WSDM '17. ACM, New York. pp 731–739. https://doi.org/10.1145/3018661.3018667

Institute for Scientific Information, Garfield E, Sher IH, Torpie RJ (1964) The Use of Citation Data in Writing the History of Science. Institute for Scientific Information. https://books.google.se/books?id=CsKXswEACAAJ

Vega *et al. Applied Network Science* (2018) 3:25

Page 25 of 26

Kanhabua N, Blanco R, Nørvåg K (2015) Temporal information retrieval. Found Trends Inf Retr 9(2):91–208. https://doi.org/10.1561/1500000043

Kim J, Diesner J (2017) Over-time measurement of triadic closure in coauthorship networks. Soc Netw Anal Min 7(1):9. https://doi.org/10.1007/s13278-017-0428-3

Kimura D, Hayakawa Y, Delvenne J-C (2008) Coevolutionary networks with homophily and heterophily. Phys Rev E 78(1):161–168. https://doi.org/10.1103/PhysRevE.78.016103

Kivelä M, Arenas A, Barthelemy M, Gleeson JP, Moreno Y, Porter MA (2014) Multilayer Networks. J Complex Netw 2(3):203–271. https://doi.org/10.1093/comnet/cnu016

Kotsakis E (2002) Structured information retrieval in xml documents. In: Proc. of the 2002 ACM Symposium on Applied Computing, SAC '02. ACM, New York. pp 663–667. https://doi.org/10.1145/508791.508919

Kralj J, Valmarska A, Grčar M, Robnik-Šikonja M, Lavrač N (2016) Analysis of Text-Enriched Heterogeneous Information Networks. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-26989-4_5

Lambiotte R, Tabourier L, Delvenne J-C (2013) Burstiness and spreading on temporal networks. Eur Phys J B 86(320). https://doi.org/10.1140/epjb/e2013-40456-9

Lambiotte R, Salnikov V, Rosvall M (2015) Effect of memory on the dynamics of random walks on networks. J Complex Netw 3(2):177–188. https://doi.org/10.1093/comnet/cnu017

Lavrenko V, Schmill M, Lawrie D, Ogilvie P, Jensen D, Allan J (2000) Mining of concurrent text and time series. In: SIGKDD workshop on text mining. ACM, New York. pp 37–44

Lee J, Zaheer M, Günnemann S, Smola A (2015) Preferential Attachment in Graphs with Affinities(Lebanon G, Vishwanathan SVN, eds.). PMLR, San Diego

Lentz HHK, Selhorst T, Sokolov IM (2013) Unfolding accessibility provides a macroscopic approach to temporal networks. Phys Rev 110(11). https://doi.org/10.1103/PhysRevLett.110.118701

Leskovec J, Krause A, Guestrin C, Faloutsos C, VanBriesen J, Glance N (2007) Cost-effective outbreak detection in networks. International conference on Knowledge Discovery and Data Mining (KDD):420. https://doi.org/10.1145/1281192.1281239

Magnani M, Rossi L (2013) Formation of multiple networks. In: International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction, Vol. 7812 of LNCS. Springer Berlin Heidelberg. pp 257–264. https://doi.org/10.1007/978-3-642-37210-0

Magnani M, Montesi D, Rossi L (2010) Friendfeed breaking news: death of a public figure. IEEE Computer Society, Los Alamitos

Magnani, M, Montesi D, Rossi L (2012) Conversation retrieval for microblogging sites. Inf Retr 15(3-4):354–372. https://doi.org/10.1007/s10791-012-9189-9

Menczer F (2004) Evolution of document networks. Proc Natl Acad Sci 101(suppl 1):5261–5265. http://arxiv.org/abs/http://www.pnas.org/content/101/suppl_1/5261.full.pdf, https://doi.org/10.1073/pnas.0307554100

Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328(5980):876–878. https://doi.org/10.1126/science.1184819

Newman MEJ (2010) Networks: An Introduction. Oxford University Press, Oxford

Nerghes A, Lee JS, Groenewegen P, Hellsten I (2014) The shifting discourse of the european central bank: Exploring structural space in semantic networks. In: 10th International Conference on Signal-Image Technology and Internet-Based Systems. pp 447–455. https://doi.org/10.1109/SITIS.2014.13

O'Connor B, Balasubramanyan R, Routledge BR, Smith NA (2010) From tweets to polls: Linking text sentiment to public opinion time series. In: Cohen WW, Gosling S (eds). Proc. of the Eleventh International Conference on Web and Social Media. The AAAI Press, Palo Alto

Paranjape A, Benson AR, Leskovec J (2017) Motifs in temporal networks. In: Proc. of the 10th ACM International Conference on Web Search and Data Mining, WSDM '17. ACM, New York. pp 601–610. https://doi.org/10.1145/3097983.3098145

Peixoto TP, Rosvall M (2017) Modelling sequences and temporal networks with dynamic community structures. Nat Commun 8(1):582–594. https://doi.org/10.1038/s41467-017-00148-9

Ren X, El-Kishky A, Wang C, Han J (2016) Automatic entity recognition and typing in massive text corpora. In: Proc. of the 25th International Conference Companion on World Wide Web, WWW '16 Companion. International WWW Conferences Steering Committee, Geneva. pp 1025–1028. https://doi.org/10.1145/2872518.2891065

Ren X, Lv Y, Wang K, Han J (2017) Comparative document analysis for large text corpora. In: Proc. of the Tenth ACM International Conference on Web Search and Data Mining, WSDM '17. ACM, New York. pp 325–334. https://doi.org/10.1145/3018661.3018690

Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents, in: Proc. of the 20th Conference on Uncertainty in Artificial Intelligence, UAI '04. AUAI Press, Arlington

Rosvall M, Esquivel AV, Lancichinetti A, West JD, Lambiotte R (2014) Memory in network flows and its effects on spreading dynamics and community detection. Nat Commun 5(1):4630–4643. https://doi.org/10.1038/ncomms5630

Roth C (2017) Knowledge Communities and Socio-Cognitive Taxonomies. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-64167-6_1

Roth C, Cointet J-P (2010) Social and semantic coevolution in knowledge networks. Soc Networks 32(1):16–29. dynamics of Social Networks. https://doi.org/10.1016/j.socnet.2009.04.005

Salehi M, Sharma R, Marzolla M, Magnani M, Siyari P, Montesi D (2015) Spreading processes in multilayer networks. IEEE Trans Netw Sci Eng 2(2):65–83. https://doi.org/10.1109/TNSE.2015.2425961

Šćepanović S, Mishkovski I, Gonçalves B, Nguyen TH, Hui P (2017) Semantic homophily in online communication: Evidence from twitter. Online Soc Netw Media 2:1–18. https://doi.org/10.1016/j.osnem.2017.06.001

Scholtes I (2011) When is a network a network?: Multi-order graphical model selection in pathways and temporal networks. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Vol. 1 of KDD 2017. ACM. pp 1037–1046. https://doi.org/10.1145/3097983.3098145

Scholtes I, Wider N, Pfitzner R, Garas A, Tessone CJ, Schweitzer F (2014) Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. Nat Commun 5(1):5024–5033. https://doi.org/10.1038/ncomms6024

Shi C, Li Y, Zhang J, Sun Y, Philip SY (2017) A survey of heterogeneous information network analysis. IEEE Trans Knowl Data Eng 29(1):17—37

Snijders TAB (2005) Models for Longitudinal Network Data, Structural Analysis in the Social Sciences. Cambridge University Press. https://doi.org/10.1017/CBO9780511811395.011

Snijders TAB (2014) Siena: Statistical Modeling of Longitudinal Network Data. Springer New York, New York. https://doi.org/10.1007/978-1-4614-6170-8_312

Sole RV, Murtra BC, Valverde S, Steels L ((2010)) Language networks: Their structure, function, and evolution. Complexity 15(6):20—26

Sowa JF (2014) Principles of semantic networks. Explorations in the representation of knowledge, Morgan Kaufmann

Tamine L, Soulier L, Ben Jabeur L, Amblard F, Hanachi C, Hubert G, Roth C (2016) Social media-based collaborative information access: Analysis of online crisis-related twitter conversations. In: Proc. of the 27th ACM Conference on Hypertext and Social Media, HT '16. ACM, New York. pp 159–168. https://doi.org/10.1145/2914586.2914589

Viard T, Latapy M, Magnien C (2016) Computing maximal cliques in link streams. Theor Comput Sci 609(1):245–252. https://doi.org/10.1016/j.tcs.2015.09.030

Wang C, Song Y, Li H, Sun Y, Zhang M, Han J (2017) Distant meta-path similarities for text-based heterogeneous information networks. In: Proc. of the 2017 ACM on Conference on Information and Knowledge Management, CIKM '17. ACM, New York. pp 1629–1638. https://doi.org/10.1145/3132847.3133029

White HD, Griffith BC (1981) Author cocitation: A literature measure of intellectual structure. J Am Soc Inf Sci 32(3):163–171. https://doi.org/10.1002/asi.4630320302

Zhou X, Hristova D, Noulas A, Mascolo C (2017) Detecting socio-economic impact of cultural investment through geo-social network analysis. In: Proc. of the Eleventh International Conference on Web and Social Media. AAAI Press, Palo Alto. pp 720–724