Evolving Networks: Eras and Turning Points

Michele Berlingerio[1], Michele Coscia[1,2], Fosca Giannotti[1],
Anna Monreale[1,2], Dino Pedreschi[2]

[1]ISTI - CNR, Area della Ricerca di Pisa, Italy {name.surname}@isti.cnr.it

[2]Computer Science Dep., University of Pisa, Italy {coscia,annam,pedre}@di.unipi.it

Abstract

Within the large body of research in complex network analysis, an important topic is the temporal evolution of networks. Existing approaches aim at analyzing the evolution on the global and the local scale, extracting properties of either the entire network or local patterns. In this paper, we focus on detecting clusters of temporal snapshots of a network, to be interpreted as *eras* of evolution. To this aim, we introduce a novel hierarchical clustering methodology, based on a dissimilarity measure (derived from the Jaccard coefficient) between two temporal snapshots of the network, able to detect the *turning points* at the beginning of the eras. We devise a framework to discover and browse the eras, either in top-down or a bottom-up fashion, supporting the exploration of the evolution at any level of temporal resolution. We show how our approach applies to real networks and null models, by detecting eras in an evolving co-authorship graph extracted from a bibliographic dataset, a collaboration graph extracted from a cinema database, and a network extracted from a database of terrorist attacks; we illustrate how the discovered temporal clustering highlights the crucial moments when the networks witnessed profound changes in their structure. Our approach is finally boosted by introducing a meaningful labeling of the obtained clusters, such as the characterizing topics of each discovered era, thus adding a semantic dimension to our analysis.

1 Introduction

One research direction in analysis of complex networks that has attracted researchers in various fields, including Data Mining, is the study of network evolution over time. Time in networks can play a double role: the entities involved may perform actions, and the connectivity structure may change. In this last setting, several phenomena can be analyzed, and much effort has been devoted in this direction so far [26, 25, 20, 5, 4].

In this paper, which extends our preliminary study [6], we focus on detecting clusters of temporal snapshots of an evolving network, to be interpreted as *eras* of evolution of the network. By analyzing the similarity of the structures of consecutive temporal snapshots of the same network, we observe that, despite global trends of similarity, it is possible to detect periods of sudden change of behavior, where people act in counter-trend, making this similarity either decrease, or suddenly start increasing very fast, much more than the average.

In many real-life social networks, in fact, a common phenomenon is that people tend to both keep being part of the networks, and keep alive all the connections created in the past. On the other hand, new users join the networks as time goes by, and people set new relationships while keeping the previous ones[26]. However, in a particular class of networks, which includes many co-authorship, transportation, and technological networks, while the number of newly created links tends to be almost constant at every snapshot, the number of previous relationships kept alive grows, thus the global effect of newly added nodes or edges looses importance over time [4]. Because of this, the similarity of the structure of two consecutive temporal snapshots increases almost at each step. The increase, however, is not locally uniform: for example, in a co-authorship network, there can be one snapshot where suddenly people change behavior and start giving more importance to creating new connections, In other words, despite a global moderate *conservative* trend, people can suddenly alternate highly more conservative periods, or a highly more *innovative* behavior.

On the other hand, there are other classes of networks that behave differently from the ones cited above, in a more *dynamic* way, where new connections replace old ones, and the importance of having new nodes or new links, globally dominates the advantages of preserving old nodes or edges. However, even in this class of evolving networks, it is possible to detect different paces of the evolution along time.

The aim in this paper is to catch these sudden changes by detecting the snapshots in which they start. Intuitively, these are starting points of new eras, i.e. *turning points* in the evolution of a network. In a globally changing world, we then want to detect eras characterized not by changes in structure (that we not only allow within the same cluster of snapshots, but we also expect), but rather characterized by a change in counter-trend with the previous era: either the previous era results more conservative, or it is actually more innovative than the era under investigation.

To this aim, we introduce a novel hierarchical clustering methodology, based on a dissimilarity measure derived from the Jaccard coefficient computed between two temporal snapshots of the network. We devise a framework to discover and browse the era hierarchy either in top-down or a bottom-up fashion, from the lowest level of the single temporal snapshots, to the highest level of the complete period of existence of the network.

In order to do so, we find a measure of the dissimilarity of two temporal snapshots, and we show how to use it as a basis for detecting starting points of new eras. We apply this methodology to three real networks, extracted by the well known bibliographic database DBLP, the movie database IMDb, and the GTD database of terrorist attacks. From these sources, we build three networks showing very different behaviors in their evolution: a co-authorship network from DBLP, a collaboration network from IMDb, and a cell-cell network from GTD, where two terrorist cells are connected if they performed an attack to the same country. On each of them, we analyze both the sets of nodes and edges, and study their evolution along time.

Our contribution can be then summarized as follows: we define a dissimilarity measure between two temporal snapshots of an evolving network, aimed at detecting turning points of the evolution; driven by this measure, we devise a methodology for hierarchically clustering the history of the network and we test our framework on real networks; we define also a methodology for adding a semantic layer to our analysis, in order to describe the eras obtained; finally, we analyze the implications of our framework to the link prediction problem.

2 Related Work

There are several studies in the literature of evolving networks. They differ by the problem treated, the level of the analysis, the solution proposed, and the networks analyzed.

Due to the difficulty of obtaining fine-grained temporal information about the arrival of a node or an edge in an evolving network, the temporal analysis of network makes often use of temporal snapshots of the evolution. In [22] authors use this method for studying the linkage pattern of blogs and the emergence of communities in the blogspace. Interesting properties have been recently studied and discovered on evolving networks, such as shrinking diameters, and densification power law. As an example, the authors in [26] discover that in most of these networks the number of edges grows superlinearly in the number of nodes over time and that the average distance between nodes often shrinks over time. In literature, many models capturing these properties have been proposed; an interesting survey is presented in [12]. In [23], Kumar et al. consider the evolution of structure within large online social networks. Specifically, they propose some measures exposing a segmentation of the networks into three regions: singletons, which do not participate in the network; isolated communities, which display star structure; and a component which persists even in the absence of stars. Three more recent works are [25, 28, 34]. In the first, Leskovec et al. present a detailed study of network evolution. They analyze four networks with temporal information about node and edge arrivals and use a methodology based on the maximum-likelihood principle to show that edge locality plays a critical role in evolution of networks. In the second, McGlohon et al. study the evolution of connected components in a network. In [34], the authors propose a novel model which captures the co-evolution of social and affiliation networks. The notion of *temporal graph* has been studied in [20]. The main aim of this paper is to study how the basic properties of graphs change over time. A similar setting is used in [21] where Kossinets et al. study the temporal dynamics of communications. They define a temporal notion of "distance" in the underlying social network measuring the minimum time required for information to spread between two nodes.

Other studies related to the temporal analysis in a network propose the study of aspects of the temporal evolution of the Web [8, 14, 17, 7]. In [8] authors study the rate of change of Web pages. In particular, they collected pages over an average interval of 37 days and base on their study on the recording of the last-modified timestamp and the downloading time of pages. Cho and Garcia-Molina in [14] propose estimators for the frequency of change of Web pages useful to improve web crawlers, web caches and to help data mining. The study presented in [17] models the persistence of both URLs and Web content finding that most URLs have a short life, while a minor fraction of pages persist for long periods of time. Bordino et al. in [7] analyze the Uk Web graph to check whether the information contained in the graph is reliable and study some aspects of the temporal evolution of this graph. In [32], Sun et al. deal with a stream of graphs, focusing on the changes within the community structure that occur over time. In order to detect changing points, their search is driver by the Minimum Description Length principle. However, their exploration can not be browsed with different temporal granularity, and they do not provide a method for the interpretation of the different eras. In [33], Tong et al. propose a method for automatically grouping time stamps into clusters as well as spot the abnormal timestamps. For each cluster/abnormal timestamp, it allows to output the selective subsets of events/entities/attribute values as their interpretations. This approach is based on a graph representation for the datasets at different timestamps and on the exploration of the proximity among different nodes. They also propose an approach for efficiently analyze multiple scales. In this case, the key idea is to explore the "smoothness" among different scales.

For our temporal analysis we perform hierarchical clustering: a survey on existing clustering approaches can be found in [3].

3 Problem Definition

We are given an evolving network G, whose evolution is described by a temporally ordered sequence of temporal snapshots $T = \{t_1, t_2, \dots, t_n\}$, where t_i represents the i-th snapshot. T can be either computed on the sets of nodes, i.e. each snapshot t_i is represented by the set of nodes involved, or on the sets of edges, i.e. each snapshot is represented by the set of edges in it.

Based on a dissimilarity measure $d : (t_i, t_{i+1}) \rightarrow [0, 1]$, we want to find a hierarchical clustering on T, returning clusters $C_i = \{t_j, \dots, t_{j+k}\}$, with $j \geq 1$, and $0 \leq k \leq n - j$.

Each cluster represents then an era of evolution. Due to the global evolution of real-life networks, we do allow alterations of the structure of the network among snapshots of the same cluster, as long as they follow a constant trend. As soon as this trend changes, we want to set the corresponding snapshot as the first of a new era, i.e. a *turning point*. The stronger is the change, the higher should be the dissimilarity of that snapshot with the previous one. The definition of the dissimilarity function should reflect this intuition.

We then want to assign to each cluster C_i a set of labels describing the represented era. This step adds a semantic dimension to our framework.

4 Framework for temporal analysis

In this section we describe the key steps composing our framework: (a) defining and computing a dissimilarity measure on the temporal snapshots; (b) merging the snapshots into clusters; (c) assigning semantics to the clusters based on frequent labels. This section provides the theoretical foundations of our framework, while next section shows the experiments performed on real networks. Both the sections are organized following the above steps.

4.1 Dissimilarity

In order to perform clustering, the first step is to define a measure of dissimilarity among elements that we want to cluster. In our setting, a simple way to do this is to use the Jaccard coefficient, to measure the correlation among the snapshots of the network. In a generic network, we can easily apply this coefficient to either two sets of nodes or two sets of edges, where each set corresponds to a temporal snapshot of the network. The coefficient would then tell us how each snapshot is correlated to the previous one, helping in detecting turning points along the evolution. As we show later in the paper, clustering temporal snapshots actually corresponds to perform a segmentation of the sequence of the snapshots, thus we are interested only in computing this Jaccard coefficient for every pair of consecutive snapshots. Note that the Jaccard coefficient could be computed between any pair of sets, thus corresponding also to non-consecutive snapshots. We are, however, not interested in a two-dimensional clustering of its values, which would lead to eras formed by potentially non-consecutive years; rather we want to perform monodimensional clustering of the temporal evolution of the Jaccard. In the experimental section we also show how, for the networks we use, this intuition is also supported by the values of the Jaccard: every snapshot is more correlated with its precedent and consecutive ones, than with any other else, justifying eras formed by consecutive snapshots.

Many real-life networks are characterized by a global evolutionary trend, then if we plot the Jaccard coefficient for each snapshot, we shall see a global trend, characterized by an almost constant slope of the Jaccard coefficient plot, alternated by (moderate to high) changes of this slope. An immediate way to define starting point of new eras is to detect the snapshots corresponding to these changes. This could be done by computing the second derivative of the Jaccard and finding values different from zero. However, the Jaccard is continuous but not derivable exactly in the points we need. To overcome this problem, we consider an approximation of the second derivative defined as follows. We take triples of consecutive years, and trace the segment that has, as endpoints, the Jaccard computed for the first and the third snapshot. If the middle point is distant from the segment, the corresponding snapshot should be considered as the start of a new era. The Euclidean distance between the middle point and the segment also gives us a quantitative analysis of how important is the change: the higher the distance, the stronger the change.

Definition 1 *Given a temporal snapshot t_j, we define the following measure:*

$$s_N(t_j) = \frac{|c_N(t_j) - (m \times j) - q|}{\sqrt{(1 + (m^2))}}$$

where $m = \frac{c_N(t_{j-1}) - c_N(t_{j+1})}{t_{j-1} - t_{j+1}}$, $q = (-(j+1) \times m) + c_N(t_{j+1})$, *and* $c_N(t_k) = \frac{|N_{k-1} \cap N_k|}{|N_{k-1} \cup N_k|}$ *is the Jaccard coefficient computed on the node sets.*

Defining s_E, which is the counterpart computed on the set of edges, requires to consider c_E instead of c_N, where c_E is the Jaccard computed on the edges.

However, this measure takes, formally, only one snapshot as input, thus it is not intuitive to use as basis for a clustering methodology. In order to tackle this problem, we define a dissimilarity between any two snapshots as follows.

Definition 2 *Given an ordered sequence $t_1, t_2, \ldots, t_n$ of temporal snapshots of a network G, the dissimilarity between any two snapshots t_i and t_j computed on their node sets is defined as*

$$d_N(t_i, t_j) = \begin{cases} s_N(t_{max(i,j)}) & if\ |i - j| = 1 \\ undefined & otherwise \end{cases}$$

Defining the similarity on the edges d_E requires to consider s_E instead of s_N. The reason of considering only subsequent snapshots is explained by looking at Figure 1: points in the timeline that are adjacent are found to be significantly more similar than distant points. We discuss further on this in Section 5.

Moreover, this dissimilarity measure allows for a straightforward hierarchical clustering: an higher dissimilarity corresponds to a stronger separation between two consecutive eras. This means that by setting a fixed threshold, we can draw a dendrogram of the hierarchical clustering, driven by this dissimilarity as a criterion for merging two consecutive clusters in a bigger one. Note that the hierarchy among clusters permits to analyze the eras with a different granularity, allowing different sensibility of the framework to the changes of the network structure.

4.2 Hierarchical clustering

Having defined a measure of dissimilarity, we are now ready to group together our snapshots into clusters, starting from single-member ones, and then merging, driven by increasing values of dissimilarity.

In hierarchical clustering, when merging clusters, there are various main approaches followed in the literature to define the distance between two clusters: the maximum distance between any two points belonging to the two clusters (complete linkage), the minimum (single linkage), the average (average linkage), the sum of all the intra-cluster variance, and so on.

Given two clusters $C_i = \{t_1, t_2, \ldots, t_k\}$ and $C_j = \{t_{k+1}, t_{k+2}, \ldots, t_{k+p}\}$, in order to define the distance between two clusters, we shall first compute all the distances between every pair (t_i, t_j), with $1 \leq i \leq k$ and $k + 1 \leq j \leq k + p$.

However, according to Definition 2, only one pair of snapshots has a dissimilarity defined: (t_k, t_{k+1}). At this point, we use this dissimilarity as inter-cluster distance. As one can see, taking the only available dissimilarity value as distance between clusters actually corresponds not only to both the complete linkage and the single linkage, but also to the average. In our case, thus, the three of them are identical.

4.3 Semantic enrichment of clusters

Once we have computed the cluster hierarchy, we want to add a description of every era. In order to do so, in analogy with the TF-IDF approach used in the Information Retrieval literature [31], we label each cluster with the node (or edge, or a property of it), that maximizes the ratio between its relative frequency in that cluster, and its relative frequency in the entire network. This strategy may produce several values equal to 1 (identical numerators and denominators). In order to discern among these cases, we weight the numerator by multiplying it again for the relative frequency in the cluster under analysis. In this way, we give more importance to 1s deriving from nodes (or edges) with a higher number of occurrences in the cluster.

With this frequency based strategy, we are assigning labels that truly characterize each cluster, as each label is particularly relevant in that cluster, but less relevant for the entire network.

One important caveat in this methodology is what to take as label for the edges. In fact, while for the nodes it is straightforward to consider the identity of the corresponding entity of the network as candidate label, the edge expresses a relationship with a semantic meaning, thus each network requires some effort in defining exactly which label could be applied to a cluster computed on edges. For example, in a co-authorship network, where two authors are connected by the papers that they have written together, a possible strategy is to take every keyword in the title of the papers as possible label. In the experimental section we show three different sets of properties used as labels for our networks.

4.4 Complexity

The entire framework requires to compute several Jaccard indexes, the dissimilarity measure and the frequencies of the labels. The computation of the Jaccard between two sets A and B requires $O(|A|+|B|)$. Thus, when computed on the sets of nodes and edges, for each network with n snapshots, we have a complexity of $O(\sum_{i=1}^{i<n}(|N_i|+|N_{i+1}|)+\sum_{i=1}^{i<n}(|E_i|+|E_{i+1}|))$, where N_i is the set of nodes of the i^{th} snapshot, and E_i is the set of edges of the i^{th} snapshot. To this, we have to add $O(2n)$ to compute the dissimilarities on both nodes and edges. We then have to add $O(n-1)$ for merging the clusters. Given W the multiset of node and edge labels, we finally have to add $O(|W|)$ to assign labels to clusters. To summarize, for each network, we have a total complexity of

$$O(\sum_{i=1}^{i<n}(|N_i|+|N_{i+1}|)+\sum_{i=1}^{i<n}(|E_i|+|E_{i+1}|)+2n+n-1+|W|)$$

$$= O(\sum_{i=1}^{i<n}(|N_i| + |N_{i+1}|) + \sum_{i=1}^{i<n}(|E_i| + |E_{i+1}|) + |W|)$$
$$= O(|N| + |E| + |W|),$$

where N is the multiset[1] of all the nodes appearing in any of the snapshots and E is the multiset of all the edges appearing in any of the snapshots, which leads to a scalable framework.

5 Experiments

As stated previously, we made use of three different sources for building our networks.

DBLP. From this bibliographic database[2], we created a co-authorship graph for the years 1955-2007, where each node represents an author and each edge a paper written together by the two connected authors. We then considered each year as temporal snapshot of DBLP, generating then 53 snapshots. In each snapshot we put only the nodes or the edges appearing in the corresponding year, thus not following a cumulative approach. The total number of resulting nodes was 582,179, with a total of 2,555,850 edges.

IMDb. From the Internet Movie Database[3], we created a collaboration graph for the years 1899-2010, where each node represents a person who took part in the realization of a movie (directors, cast, song writers, and so on), and two persons are connected if they participated to the realization of the same movie. We considered each year as temporal snapshot, generating then 112 snapshots. As for DBLP, the snapshots are non-cumulative, for a total number of 57,457 nodes and 13,047,319 edges.

GTD. From this database of global terrorism[4], we created a group-group graph for the years 1969-2008, where each node represents a terrorist group or organization, and two groups are connected if they participated in a terrorist attack to the same country (note that the two groups only attacked the same country, but they do not need to have collaborated to the attack in order to be connected). We then considered each year as temporal snapshot, generating 40 snapshots. As for DBLP, the snapshots are non-cumulative, and we ended up with 2,279 nodes and 31,843 total edges.

For each of the networks we also built synthetic null models reflecting the global statistics of the network, in terms of number of snapshots and number of edges per snapshot. We created two different null models for each network:

Random. Nodes and edges are placed at random, only the statistics of the original networks were preserved.

Preferential attachment. While preserving the number of snapshots and the number of edges per snapshot, each snapshot is created following the preferential attachment model[2], i.e. the probability of connecting two nodes is directly

[1] We have multisets because every node or edge can be found in more than one snapshot

[2] http://dblp.uni-trier.de

[3] http://www.imdb.com

[4] http://www.start.umd.edu/gtd

proportional to their degrees. Please note that the generator is ran separately for each snapshot.

All the experiments where performed on a server equipped with a dual Xeon @ 3.06Ghz, 8GB of Ram, running the Ubuntu 8.04 Server 64bit operating system. In line with the theoretical complexity shown in the previous section, each network required less than ten minutes of total computation, and less than 500 megabytes of ram to be processed, despite the size of the networks.

5.1 Jaccard Coefficient

Figures 2(a,c,e) show the Jaccard computed on both the node and the edge sets. These plots report a general increasing behavior of the Jaccard during time in DBLP, both on nodes and on edges, broken by short series of years in which people acted in counter-trend. On the other hand, for the other two networks the temporal behavior seems not to follow a specific trend, while, in particular, GTD presents a hole of two years in the history of the network.

Two questions might be raised on the effectiveness of following a Jaccard-based approach for clustering eras: what would the Jaccard computed on non consecutive snapshot tell us? Are we dealing with some random or real phenomena?

We start answering the first question by plotting the coefficient computed for every pair of snapshots: Figure 1 shows that the Jaccard decreases when computed between snapshots more distant in time. As stated in the previous section, this observation justifies a dissimilarity measure that takes into account only consecutive snapshots, as two distant snapshots are not likely to be similar, thus they will belong to different clusters. Temporal segmentation is then a good model for clustering real-life evolving networks, which is a consideration well accepted in the literature regarding evolving networks [4, 26].

Answering the second question requires to compare the knowledge extracted with our methodology on real and random networks. If such knowledge is similar, we might conclude that our methodology is not able to extract any useful, non-random, information. We then followed an approach which is common in the network analysis literature [15]: building random networks as null models and testing the framework on them. In order to do so, we created random and preferential attachment null models, as stated at the beginning of this section, and computed the Jaccard on them. As we see in figures 2(a,c,e), the random component of both the null models makes the framework not meaningful on them. Note that the lines corresponding to the random and the preferential attachment networks follow the zero constant. This might be surprising for the preferential attachment, but we recall that the generator was reset at every snapshot, thus making two consecutive snapshots randomly correlated.

5.2 Dissimilarity

The second step of the framework requires to compute our dissimilarity on the basis of the Jaccard coefficient computed on the network. Figures 2(b,d,f) report the values of the dissimilarity for both the edges and the nodes, for each network. As one can see, the quantitative analysis of our dissimilarity measure is effective: its values have a considerable standard deviation. That is, we can effectively perform hierarchical clustering finding a well distributed strength of starting snapshots for new eras of evolution.

Another observation that can be done is that while the Jaccard values computed on nodes or edges show similar trends, stronger differences can be found in the dissimilarity plots. That is, we expect the eras computed on nodes slightly differ from the ones computed on the edges.

As last note, we see that in the first years, although not always noticeable from the Jaccard plots, the dissimilarity spots very unstable behavior. This could be mainly explained by two considerations: first, at the beginning of the history of every network, the network structure is still very little, and a change of a few nodes or edges may result in a strong change of the Jaccard values; second, even though a network follows one clear model of evolution (DBLP is well known to follow the preferential attachment model [4, 10]), the model itself takes a few years to warm up and to be fully functional (note that in the preferential attachment this means that nodes are still not affected by the aging effects).

5.3 Hierarchical clustering

We then started to compute the clusters on the sequences of temporal snapshots. We started from clusters containing only one year and then, driven by the dissimilarity values computed in the previous step, we merged similar consecutive clusters, with increasing values of dissimilarity.

Figures 3,4,5,6 report all the dendrograms of the extracted eras for each network. Note that, due to the large number of snapshots and to the wide range of values taken by the dissimilarity, we could not plot the dendrograms with the height proportional to the dissimilarity values itself.

A few considerations can be done by looking at the dendrograms. First, in all the three networks, as expected by looking at the dissimilarity plots, the first years tend to form eras by themselves, and this is true both for nodes and for edges.

Second, while, as we said above, the Jaccard plots of nodes and edges for each network tend to look similar, and the differences are then emphasized in the dissimilarity plots, by looking at the shape of the dendrograms, discerning between eras that coincides for both nodes and edges, and eras that include different years for the two sets, appears to be easier. For example, look at the years 2001-2006 in both DBLP nodes and edges: it is easier to see those years grouped in the same era in the dendrograms in Figure 3 than in the dissimilarity plot in Figure 2(b). Same discussion for the eras 1995-1997 and 1998-2007 in the GTD network, that are both similar when comparing nodes and edges, but for which the dissimilarity plot does not clearly reflect this situation.

Third, while the dendrograms can spot situations as above, they can also highlight differences in the node and edge evolutions. Take, for example, years 1930-2009 in both nodes and edges in IMDb, as reported in Figure 4 and 5.

This era presents very different sub-eras when looking at nodes or edges, and this is because of the different importance, given during time, to new nodes or new edges over old ones.

5.4 Semantic enrichment of clusters

As last step in our framework, we computed the labels for each cluster obtained. We recall that for each cluster C_i we assign the set of the k labels maximizing the ratio between their frequency in C_i and their frequency in the entire network. Tables 1, 2 and 3 report a few of the most characterizing labels for some of the eras of each network. Due to the impressive number of total eras and labels, we could not report all the labels for all the eras, but instead we chose, for each network a selection of interesting eras (covering the entire network history), and a selection of the most representative labels for them. The DBLP keywords were pre-processed using the Porter's stemming algorithm [30].

We chose some relatively small eras in order to cover approximately the entire time span of the dataset. The start and end years of an era were selected where the inclusion of the following, or preceding, year would have caused the merging of two eras resulting in a selected period of many years not strongly correlated each other, according to the dendrogram. Verifying the labels of the extracted eras provides two benefits: it is useful to evaluate our results, as we refer to fields in which there is a ground truth about periods, and may lead to novel points of view about the history of our data sources.

For DBLP, we present the labels for the node and edge eras in Table 1. It is possible to spot some interesting eras, such as the ALGOL era from 1963 (the year of one major revision of ALGOL60[5]) to 1970. In the 70s many popular programming languages were developed, such as C (developed from 1969 to 1973[6]), Prolog (which was born in 1972 from a project aimed not at producing a programming language but at processing natural languages [16]) and Pascal (standardized in 1983[7], and this might explain also its era from 1980 to 1982).

Interesting enough, from 2004 we are witnessing a brand new era, made of networks and the increasing complexity of web technologies. Node era labels for DBLP let emerge some other key research results: we can recognize the huge work made by David Chaum (1985-1991) in the field of cryptography, the basis of the electronic currency system, culminating in 1990 with the foundation of his electronic cash company; another example is Raymond F. Boyce, a key researcher for the development of SQL [13], died in 1974.

For IMDb, we present the era labels in Table 2. It is possible to perform an analysis at two different granularity levels. At a high level, one may notice that the keywords for periods before 1975 are very specific and referring to precise concepts in movie history (such as the sound synchronized to record, referring to the very first movies with sound, or heimatfilm, such as “Lassie come home”), while after 1975 keywords are simpler and less specific (love, death,

[5] http://www.masswerk.at/algol60/report.htm

[6] http://cm.bell-labs.com/cm/cs/who/dmr/chist.html

[7] ISO 7185, http://www.pascal-central.com/iso7185.html

murder, blood). This is due to the fact that the keywords are user-assigned, thus very old movies are only watched (and tagged) by a niche of expert cinephiles, while the mass tags recent blockbusters. Note also that the vast majority of IMDb users are Western and particularly American, thus the keywords are heavily unbalanced on Hollywood and European industry, disgreaging other filmographies such as Japan, Hong Kong and the prolific Bollywood. At a lower level of granularity, our technique is able to spot actual eras or sub-eras of movie history, such as the "pre code" era (from 1930, the year in which the Motion Picture Production Code was written, to 1933, when the code become effectively enforced[8]).

In IMDb node eras we see the most prolific people in movie industry. Especially in latter years, counter-intuitively, instead of finding movie stars, which are involved in leading roles in big productions (thus it is impossible for them to participate to more than 4-5 movies a year), we see actors that are prolific in minor roles, or producers (Andreas Schmid, producer from 2004 of movies like "The Punisher", "Lord of war" and "Perfume: The Story of a Murderer", before stopping his career in 2007[9]), directors (Peter Elfelt, very well known for many experimental documentary shorts until 1907[10]) and composers. Exceptions to this rule are the extremely prolific Indian stars like Brahmanandam[11], or Hong Kong superstar Tony Leung Ka Fai, who between 1991 and 1995 appeared in many movies of the most important Hong Kong authors such as Tsui Hark, Gordon Chan and Wong Kar Wai.

Finally, consider the eras emerging in GTD dataset, for which we report the labels in Table 3. It is interesting to note that the 1977-1983 edge era was dominated by European countries, particularly Italy and France. This period coincides with the years of activity of the Hyperion School, founded in 1976 in Paris and whose members were arrested in 1983. Hyperion is considered linked with many terroristic cells in all Europe, particularly Italy[12], whose activities culminated in 1978 with the kidnapping and assassination of Italian prime minister Aldo Moro by Red Brigades. Also the node era from 1978 to 1981 witnessed the terror war fought in Italy in this period, by two extremist groups of opposite philosophy: the Marxist-Leninist group Prima Linea and the neofascist group Armed Revolutionary Nuclei (NAR). NAR was responsible, among others, of the 1980 bombing of the Bologna main train station[13]; Prima Linea had carried 18 out of their 23 assassinations from 1978 to 1981[14].

It is interesting to note that, among the sets of labels found to be character-

[8] Mick LaSalle, "Complicated Women: Sex and Power in Pre-Code Hollywood"

[9] http://www.imdb.com/name/nm1209077/

[10] http://www.imdb.com/name/nm0253298/

[11] http://www.imdb.com/name/nm0103977/

[12] Antonio Ferrari, "In teleselezione dalla Francia gli ordini ai terroristi italiani?", Corriere della Sera 26 aprile 1979

[13] 85 victims, ref. Davies, Peter, Jackson, Paul (2008). "The far right in Europe: an encyclopedia". Greenwood World Press, p. 238

[14] Presidenza della Repubblica, "Per le vittime del terrorismo nellItalia repubblicana: giorno della memoria dedicato alle vittime del terrorismo e delle stragi di tale matrice", 9 maggio 2008 (Rome: Istituto poligrafico e Zecca dello Stato, 2008, ISBN 978-88-240-2868-4)

istic for an era, there are only a few of them which were somehow “popular”. This might seem a problem of the methodology, but it is essentially due to the frequency-based approach. In the future, we plan to investigate the possibility of comparing several different alternatives, based, perhaps, on PageRank, Hits, and other measures of centrality.

6 Turning points and link prediction

As we said from the beginning, in our problem we do allow evolution within one specific era, while two subsequent eras are characterized by different paces at which the evolution takes place. Building up a model of network evolution is the task at the basis for link prediction, i.e., the problem of deciding, with a certain score, whether two nodes will link in the future [27]. There are several studies regarding link prediction, and most of them rely on a underlying model of network evolution [24, 1, 2, 18, 9, 26, 29, 19]. However, not all the models fit all the different types of networks, and most predictors perform well on certain networks, but relatively bad on others. To cope with this, recently the authors of [4, 10] introduced a supervised approach based on extracting *graph evolution rules*, i.e., local frequent subgraphs expressing evolution. The model of evolution itself is there learned from the data, by means of the extraction of those rules, that are afterward used to predict the evolution of the network. In contrast with the previous approaches, this approach allows to predict also *when* then new links will form.

However, to the best of our knowledge, all of the current approaches assume that the model of evolution is static, i.e., there is one rule (Jaccard, Common Neighbors, Adamic-Adar, Forest Fire, and so on) or a set of them (the complete set of rules extracted by GERM), governing the creation of new links, that do not change over time.

This is in contrast with our framework, where we detect moments along the evolution of a network in which the underlying evolution rule changes pace. According to this, we could state that the arrive of a new, sudden, turning point, may invalidate future predictions, as the evolution would change pace.

A question then arises: can we somehow forecast the arrive of a new era? The answer would probably change the way we currently see the link prediction problem, for the reasons we stated above.

While link and evolution prediction are not the main focus of this paper, in this section we would like to pose the basis for future work in which we will be trying to solve the link prediction problem taking eras into account. In this section, instead, we try to answer two different questions: do the temporal series formed by our dissimilarities follow any pattern? Is there a way to forecast the subsequent values of the dissimilarities?

In the rest of this section we try to answer the above questions, by means of statistical analysis of time series, and, in particular, by means of autoregressive models.

6.1 Time series analysis by autoregressive models

An autoregressive model (AR) is a type of random process often used to forecast future values of time series representing natural and social phenomena. The notation $AR(p)$ refers to the autoregressive model of order p, defined as

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, c is a constant and ε_t is white noise. Many authors omit the constant for simplicity.

We refer to [11] for a complete introduction to time series analysis, and how to perform prediction on them based on autoregressive models. We used the *tseries* package under the *R* statistical software[15] to fit autoregressive models on our dissimilarity series, and to perform prediction on them.

6.2 Forecasting dissimilarities

Figure 7 reports, for each network, five new values of dissimilarities forecast both on the edges and the nodes. These values were obtained by fitting autoregressive models as explained above, and then using the fits to forecast subsequent values. What we see in the figure is that, while in IMDb the model is forecasting relatively low values of the dissimilarities, this is not true for the other two networks, and in particular for GTD. The intuition behind these plots is that, if the forecast values are low, we are not expecting a sudden change of era in the near future, i.e., a link predictor trained on the past evolution of the network, may perform well for the near future. On the other hand, in networks where the forecast values are high, as in GTD - particularly for the nodes -, we do expect a new, well distinct, era in the next few years, with this meaning that the results of a link predictor based on the previous history of the network may be not accurate, due to the expected change of evolution pace.

The above might suggest a new way of looking at the link prediction problem, where the basic rules of evolution are supported by a certain confidence in the prediction given also by our temporal analysis of the network evolution. In the future, we plan to investigate the possibility of building such solution for the link prediction problem, based on our clustering framework.

7 Conclusions and Future Work

We have proposed a framework for the discovery of eras in an evolving social network. Based on a dissimilarity measure derived from the Jaccard coefficient, we have presented a methodology to perform hierarchical clustering of the temporal snapshots of a network. We have applied our methodology to real-life data and null models, showing the effectiveness of our approach. The semantic layer provided by the cluster labeling allows also to give an interpretation of the eras found. We believe that our work completes the wide literature in the analysis of

[15] http://www.r-project.org

evolving networks, and raises questions that do not have received considerable attention so far, providing also a possible way of answering them.

Our methodology can also put the link prediction problem under a different light, and we believe that building a predictor that takes into account also the history of eras of the network based on our findings deserves further effort and is the basis for future work. Another future research direction is to compare the results obtained by our labeling methodology, with different measures of centrality of the labels into the network, to try to explain the relationships between frequent labels (as we have), with "famous" labels (actors, researchers, and so on) that one might expect. Finally, one consideration regarding the evaluation of clusters. One typical question in clustering is how to find the optimal cut of the dendrogram. In our work, this translates into finding a measures to evaluate the best temporal granularity for our analysis. To this purpose, we plan to investigate the possibility of finding the equivalent of the *modularity* measure for the Community Discovery problem, to be seen as a measure returning the most informative temporal granularity of analysis.

8 Acknowledgments

Michele Coscia is a recipient of the Google Europe Fellowship in Social Computing, and this research is supported in part by this Google Fellowship.
We are also grateful to Giulio Rossetti for his help with the GTD database.

References

[1] A. Barabasi, H. Jeong, Z. Neda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. http://arxiv.org/abs/cond-mat/0104162, 2001.

[2] A.L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[3] Pavel Berkhin. Survey of clustering data mining techniques. Technical report, Accrue Software, San Jose, CA, 2002.

[4] Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. Mining graph evolution rules. In *ECML/PKDD (1)*, pages 115–130, 2009.

[5] Michele Berlingerio, Michele Coscia, and Fosca Giannotti. Mining the temporal dimension of the information propagation. In *IDA*, pages 237–248, 2009.

[6] Michele Berlingerio, Michele Coscia, Fosca Giannotti, Anna Monreale, and Dino Pedreschi. As time goes by: Discovering eras in evolving social networks. In *PAKDD (1)*, volume 6118, pages 81–90. Springer, 2010.

[7] Ilaria Bordino, Paolo Boldi, Debora Donato, Massimo Santini, and Sebastiano Vigna. Temporal evolution of the uk web. In *ICDM Workshops*, pages 909–918, 2008.

[8] Brian E. Brewington and George Cybenko. Keeping up with the changing web. *IEEE Computer*, 33(5), 2000.

[9] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the 7th International Conference on World Wide Web*, 1998.

[10] Björn Bringmann, Michele Berlingerio, Francesco Bonchi, and Aristides Gionis. Learning and predicting the evolution of social networks. *IEEE Intelligent Systems*, 25(4):26–35, 2010.

[11] P. J. Brockwell, R. A. Davis, and I. Netlibrary. Introduction to time series and forecasting. 2002.

[12] D. Chakrabarti and C. Faloutsos. Graph mining: Laws, generators, and algorithms. *ACM Comput. Surv.*, 38(1), 2006.

[13] Donald D. Chamberlin and Raymond F. Boyce. Sequel: A structured english query language. In Randall Rustin, editor, *SIGMOD Workshop, Vol. 1*, pages 249–264. ACM, 1974.

[14] Junghoo Cho and Hector Garcia-Molina. Estimating frequency of change. *ACM Trans. Internet Techn.*, 3(3):256–290, 2003.

[15] Aaron Clauset, Cristopher Moore, and M. E. J. Newman. Structural inference of hierarchies in networks. *CoRR*, abs/physics/0610051, 2006.

[16] Alain Colmerauer and Philippe Roussel. The birth of Prolog. In *The second ACM SIGPLAN conference on History of programming languages*, pages 37–52. ACM Press, 1993.

[17] Daniel Gomes and Mário J. Silva. Modelling information persistence on the web. In *ICWE '06: Proceedings of the 6th international conference on Web engineering*, pages 193–200, 2006.

[18] G Jeh and J Widom. Simrank: a measure of structural-context similarity. In *in KDD '02: Proceedings of the eighth ACM SIGKDD international.* ACM Press, 2002.

[19] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, (18):39–43, 1953.

[20] David Kempe, Jon M. Kleinberg, and Amit Kumar. Connectivity and inference problems for temporal networks. In *STOC*, pages 504–513, 2000.

[21] Gueorgi Kossinets, Jon M. Kleinberg, and Duncan J. Watts. The structure of information pathways in a social communication network. In *KDD*, pages 435–443, 2008.

[22] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. *World Wide Web*, 8(2):159–178, 2005.

[23] Ravi Kumar, Jasmine Novak, and Andrew Tomkins. Structure and evolution of online social networks. KDD '06, pages 611–617, New York, NY, USA, 2006. ACM.

[24] Lada. Adamic and eytan adar. friends and neighbors on the web. *Social Networks*, (25):230, 2001.

[25] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD*, pages 462–470, 2008.

[26] Jure Leskovec, Jon M. Kleinberg, and Christos Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD*, pages 177–187, 2005.

[27] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. CIKM '03, pages 556–559, New York, NY, USA, 2003. ACM.

[28] M. McGlohon, L. Akoglu, and C. Faloutsos. Weighted graphs and disconnected components: patterns and a generator. In *KDD*, pages 524–532, 2008.

[29] Michael Mitzenmacher. A brief history of lognormal and power law distributions. *Internet Mathematics*, (1):2004.

[30] Stephen E. Robertson, C. J. van Rijsbergen, and Martin F. Porter. Probabilistic models of indexing and searching. In *SIGIR*, pages 35–56, 1980.

[31] Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management*, pages 513–523, 1988.

[32] Jimeng Sun, Christos Faloutsos, Spiros Papadimitriou, and Philip S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In Pavel Berkhin, Rich Caruana, and Xindong Wu, editors, *KDD*, pages 687–696. ACM, 2007.

[33] Hanghang Tong, Yasushi Sakurai, Tina Eliassi-Rad, and Christos Faloutsos. Fast mining of complex time-stamped events. In James G. Shanahan, Sihem Amer-Yahia, Ioana Manolescu, Yi Zhang, David A. Evans, Aleksander Kolcz, Key-Sun Choi, and Abdur Chowdhury, editors, *CIKM*, pages 759–768. ACM, 2008.

[34] Elena Zheleva, Hossam Sharara, and Lise Getoor. Co-evolution of social and affiliation networks. In *KDD*, pages 1007–1016, 2009.

DBLP - Edge labels		
Start	End	Labels
1956	1962	tunnel diode, q-d-algorithm, megabits-sec, four megacycles, bounded transition
1963	1970	prediscuss, algol, machine to man, ssdl, tree manipulation
1971	1973	lr0, word functional, optimal, virtualize, syntax analysis
1975	1979	data, language, program, computer, codasyl
1980	1982	pascal, language, database, data, micro-computer
1983	1985	prolog, database, online, abstract, expert
1987	1991	parallel, program, logic, abstract, database
1992	1996	parallel, program, logic, object oriented, computer
1997	1999	model, parallel, design, distributed, image
2001	2003	model, data, network, design, image
2004	2005	network, model, algorithm, web, data

DBLP - Node labels		
Start	End	Labels
1957	1959	Yu. A. Shreider, I. Y. Akushsky, Howard H. Aiken, D. G. Hays, W. L. van der Poel
1960	1963	Calvin C. Elgot, W. D. Frazer, Roger E. Levien, Robert O. Winder, Lorenzo Calabi
1964	1972	R. L. Beurle, Sheila A. Greibach, Rina S. Cohen, Karl K. Pingle, James L. Parker
1973	1976	Raymond F. Boyce, Michael Ian Shamos, Matthew M. Geller, Louis Pouzin, Irving L. Traiger
1977	1982	Peter Raulefs, Gary G. Hendrix, Helmut K. Berg, Nathan Goodman, S. Bing Yao
1983	1984	Hans Bekic, Gunter Spur, Werner Frey, Frank-Lothar Krause, Ashok K. Thareja
1985	1991	Walter Ameling, Ehud Y. Shapiro, David Chaum, Setrag Khoshafian, David W. Stemple
1992	1996	Robert K. Brayton, Alberto L. Sangiovanni-Vincentelli, Terence C. Fogarty, Janak H. Patel, Martin Kummer
1997	2000	Miodrag Potkonjak, Bruce Schneier, Christopher J. Taylor, Alok N. Choudhary, Prithviraj Banerjee
2001	2006	Mahmut T. Kandemir, Zhaohui Wu, HongJiang Zhang, Wei-Ying Ma, Wen Gao

Table 1: Era labels on both DBLP edges and nodes

IMDb - Edge labels		
Start	End	Labels
1900	1907	spanish-american-war, early-sound, america's-cup, synchronized-to-record, trick-film
1908	1909	synchronized-to-record, film-d'art, william-shakespeare, early-sound, te-deum
1910	1912	trick-photography, broncho-billy, animal-actor, melodrama, law-enforcer
1913	1915	broncho-billy, mister-jarr, universal-ike-series, americana, ham-and-bud-series
1917	1929	melodrama, society, mutt-and-jeff, fable, world-war-one
1930	1933	pre-code, bimbo-the-dog, talkartoon, flip-the-frog, two-reeler
1935	1941	1930s, gunfire, b-movie, beautiful-woman, stock-footage
1942	1954	beautiful-woman, 1940s, usa, world-war-two, series
1956	1957	beautiful-woman, heimatfilm, 1950s, mr-magoo, sportscope
1958	1963	peplum, loopy-de-loop, modern-madcaps, independent-film, nudie-cutie
1964	1965	swifty-and-shorty, beautiful-woman, independent-film, nudie-cutie, peplum
1966	1972	female-nudity, independent-film, spaghetti-western, beautiful-woman, hippie
1973	1974	female-nudity, blaxploitation, hoot-kloot, grindhouse, martial-arts
1975	1977	independent-film, erotic-70s, poliziottesco, italian-sex-comedy, naziploitation
1979	1989	nudity, cult-favorite, murder, electronic-music-score, violence
1990	1993	murder, sequel, male-female-relationship, family-relationships, police
1994	1999	independent-film, female-nudity, gay-interest, love, friendship
2000	2002	independent-film, gay-interest, friendship, female-nudity, flashback
2004	2008	love, death, independent-film, blood, family-relationships

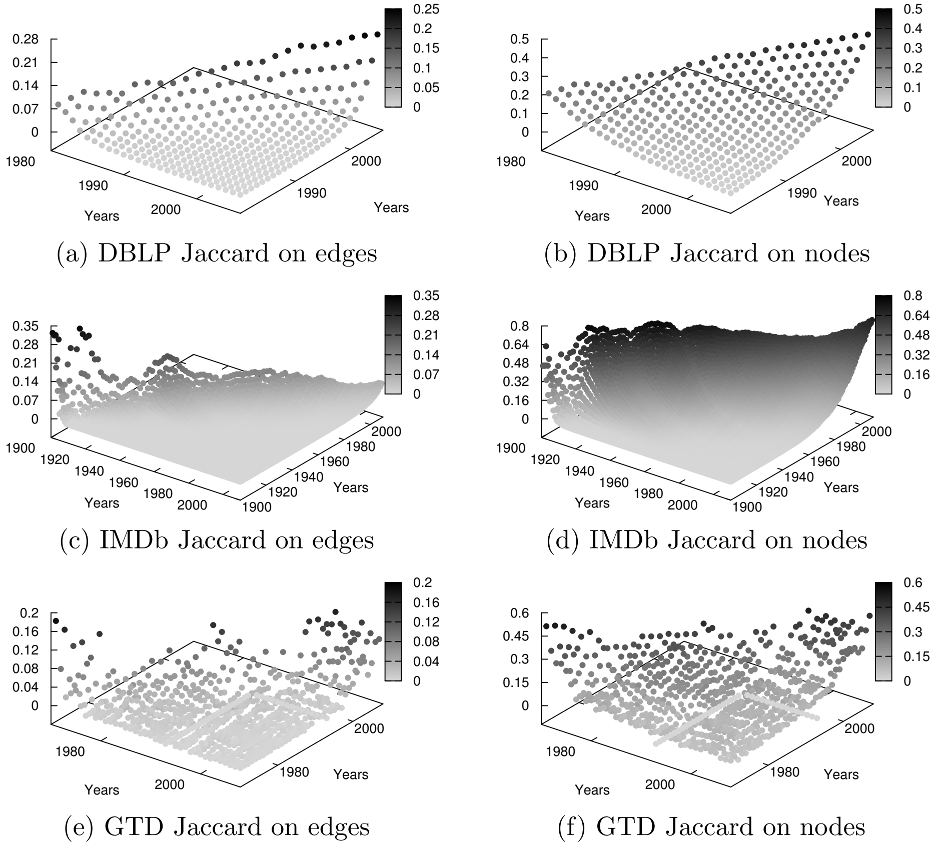
IMDb - Node labels		
Start	End	Labels
1902	1907	Alf Collins, Peter Elfelt, Lucien Nonguet, Arthur Gilbert, Alice Guy
1909	1915	Siegmund Lubin, Arturo Ambrosio, William Nicholas Selig, Pat Powers, David Horsley
1916	1922	John Randolph Bray, Matsunosuke Onoe, Burton Holmes, Bud Fisher, William Randolph Hearst
1923	1929	Abe Stern, Julius Stern, Jack White, Hal Roach, Paul Terry
1930	1931	Arthur Hurley, Leroy Shield, James Mulhauser, Amadee J. Van Beuren, Albert H. Kelley
1932	1938	Edward LeSaint, Earl Dwire, Dennis O'Keefe, Harry Bowen, Fred Parker
1939	1946	John Tyrrell, Emmett Vogan, Cyril Ring, Jack Gardner, John Dilson
1947	1952	Sam Buchwald, Edward Selzer, Stanley Wilson, Izzy Sparber, Marshall Reed
1953	1963	Milt Franklyn, Ahmet Tarik Teke, Nicholas Balla, Seymour Kneitel, Julian Biggs
1966	1975	Sung-il Shin, David H. DePatie, Luigi Antonio Guerra, Adoor Bhasi, Jeong-geun Jeon
1976	1980	Richard Lemieuvre, Cyril Val, Dominique Aveline, John Seeman, Peter Katadotis
1981	1984	George Payne, Herschel Savage, Ilayaraja, Mona Fong, Paul Thomas
1985	1990	Amrish Puri, Lily Y. Monteverde, Yunus Parvez, Shui-Fan Fung, Tony Fajardo
1991	1996	Brahmanandam, Ilayaraja, Floyd Elliott, Milind Chitragupth, Tony Leung Ka Fai
1997	2003	Brahmanandam, Johnny Lever, Phil Hawn, Yiu-Cheung Lai, Simon Lui
2004	2007	Venu Madhav, Brahmanandam, Himesh Reshammiya, Andreas Schmid, Suneel
2008	2009	Kevin MacLeod, Jose Rosete, Suraaj Venjarammoodu, Brian Jerin, Moby

Table 2: Era labels on both IMDb edges and nodes

GTD - Edge labels		
Start	End	Labels
1971	1975	United States, Northern Ireland, West Germany (FRG), France, Argentina
1977	1983	Italy, France, Spain, El Salvador, Guatemala
1984	1988	Lebanon, Colombia, Sri Lanka, France, Peru
1989	1991	India, Colombia, Israel, Myanmar, Lebanon
1992	1994	India, Bangladesh, Germany, West Bank and Gaza Strip, Venezuela
1995	1997	India, Bangladesh, Pakistan, Indonesia, Colombia
1998	1999	Greece, India, Timor-Leste, Northern Ireland, Kosovo
2000	2002	India, West Bank and Gaza Strip, Israel, Russia, Macedonia
2003	2005	Iraq, India, West Bank and Gaza Strip, Saudi Arabia, Pakistan
2006	2007	Iraq, India, Pakistan, Nigeria, Sudan

GTD - Node labels		
Start	End	Labels
1971	1975	Black September, National Front for the Liberation of Cuba (FLNC), Weatherman, Secret Cuban Government, National Integration Front(FIN)
1976	1977	Communist Combat Unit, Armed Communist Struggle, Baader-Meinhof Group, Black Order, Che Guevara Brigade
1978	1981	Armenian Secret Army for the Liberation of Armenia, Armed Revolutionary Nuclei (NAR), Right-Wing Extremists, Spanish Basque Battalion (BBE), Prima Linea
1982	1987	Armenian Secret Army for the Liberation of Armenia, Abu Nidal Organization (ANO), Anti-terrorist Liberation Group (GAL), M-19 (Movement of April 19), Action Directe
1989	1991	Moslem Janbaz Force, Bhinderanwale Tiger Force of Khalistan (BTHK), Popular Militia (Colombia), Kurdish Dissidents, Death to Bazuqueros
1992	1993	Khasi Students Union, Jharkhand Tribal Forces, Revolutionary Security Apparatus, Allah's Tigers, Ikhwan-ul-Muslimeen
1995	1997	Kuki tribesmen, Jammu and Kashmir Islamic Front, Harkat ul Ansar, Tamil Nadu Liberation Arm, Al Faran
1998	2000	Communist Party of India Marxist-Leninist, Vishwa Hindu Parishad (VHP), Individual, Shiv Sena, Mazdoor Kisan Sangram Samiti (MKSS)
2002	2005	Al-Mansoorian, Kuki Revolutionary Army (KRA), Jaish-e-Mohammad (JeM), Rashtriya Swayamsevak Sangh, Tawhid and Jihad

Table 3: Era labels on both GTD edges and nodes



(a) DBLP Jaccard on edges

(b) DBLP Jaccard on nodes

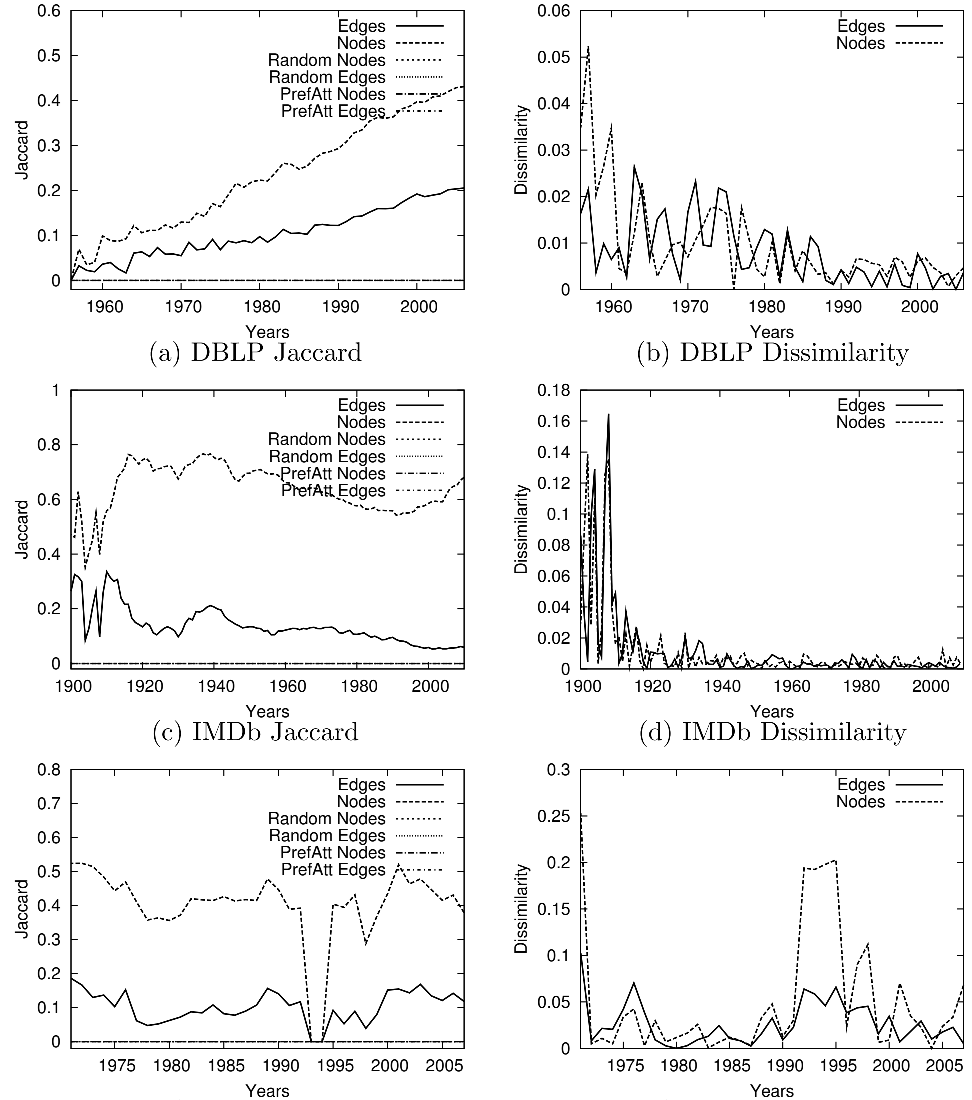
(c) IMDb Jaccard on edges

(d) IMDb Jaccard on nodes

(e) GTD Jaccard on edges

(f) GTD Jaccard on nodes

Figure 1: The Jaccard correlation computed among all the snapshots, on both edges and nodes



(a) DBLP Jaccard

(b) DBLP Dissimilarity

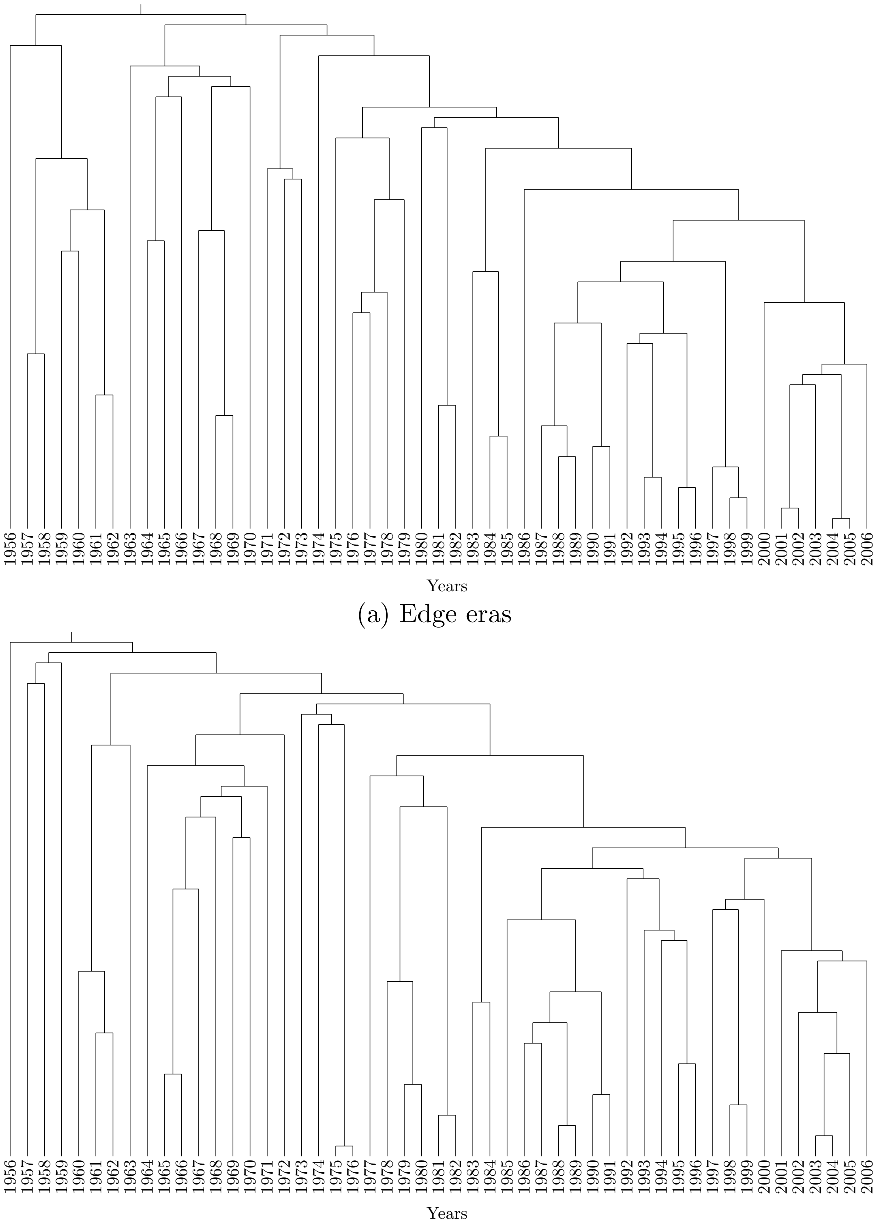
(c) IMDb Jaccard

(d) IMDb Dissimilarity

(e) GTD Jaccard

(f) GTD Dissimilarity

Figure 2: The Jaccard correlation computed only between subsequent snapshots (a,c,e) and the corresponding dissimilarities computed on it (b,d,f)



(a) Edge eras

(b) Node eras

Figure 3: Eras on both edge and node evolutions in DBLP

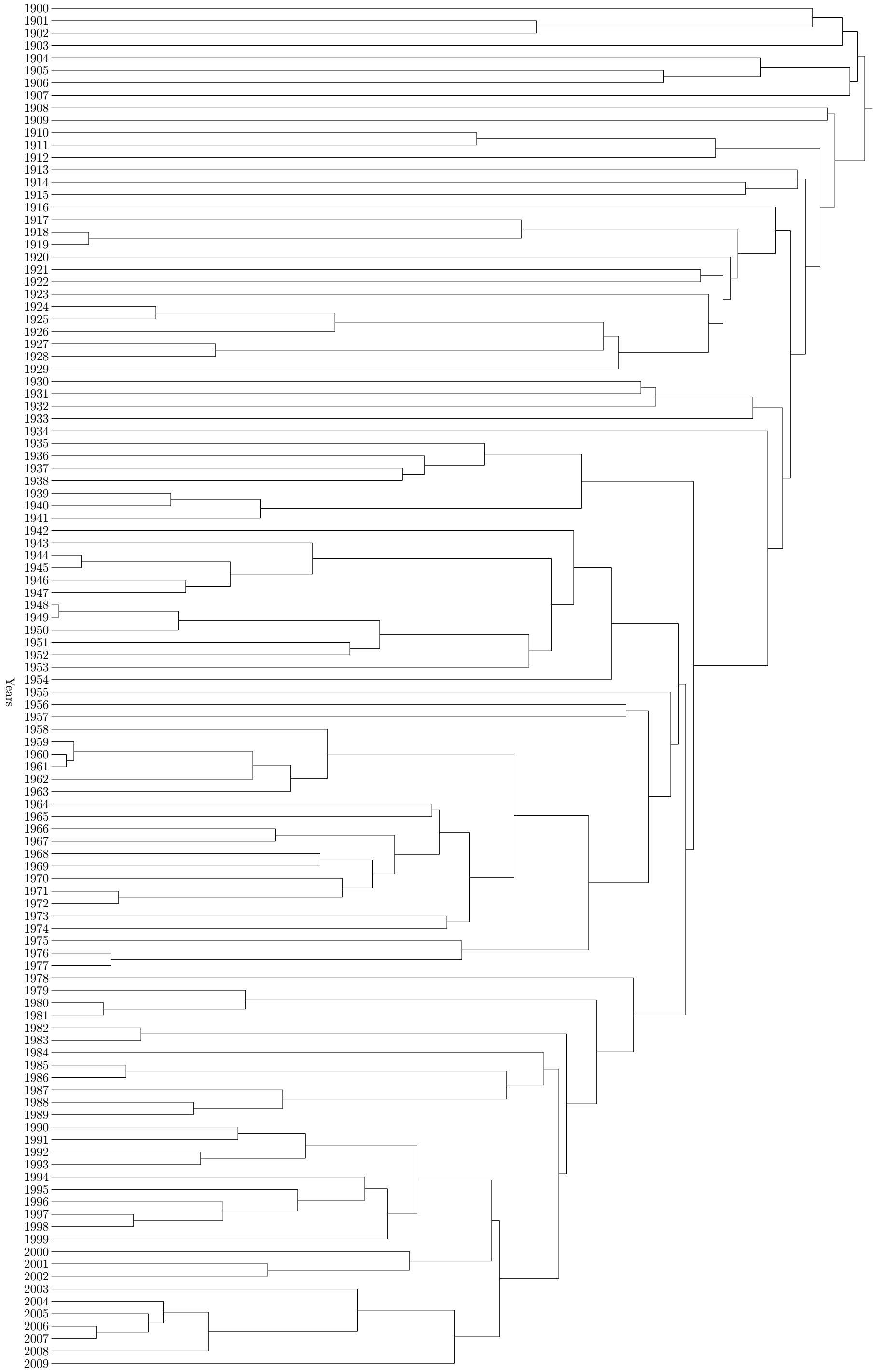


Figure 4: Eras in IMDb edge evolution

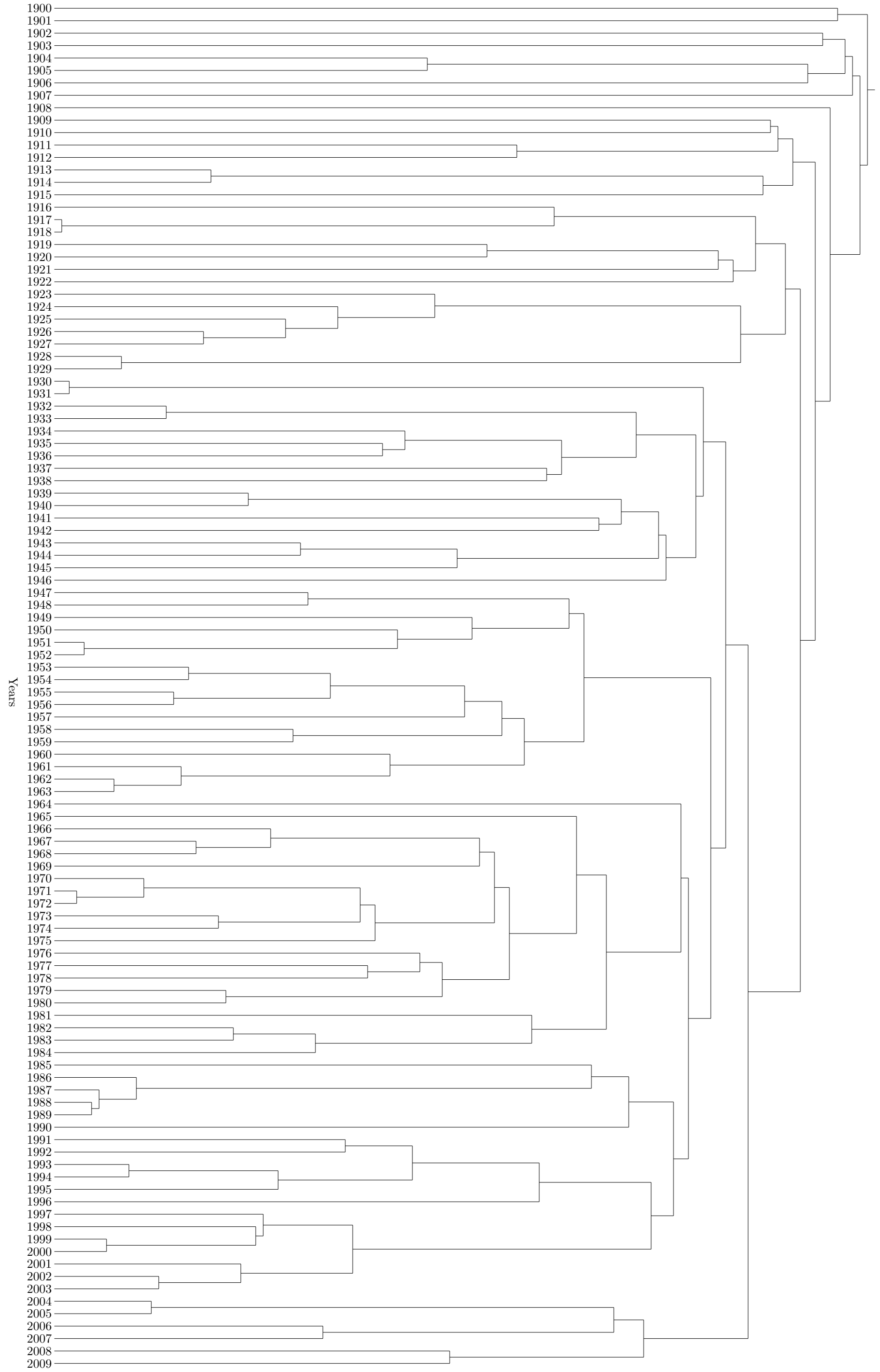
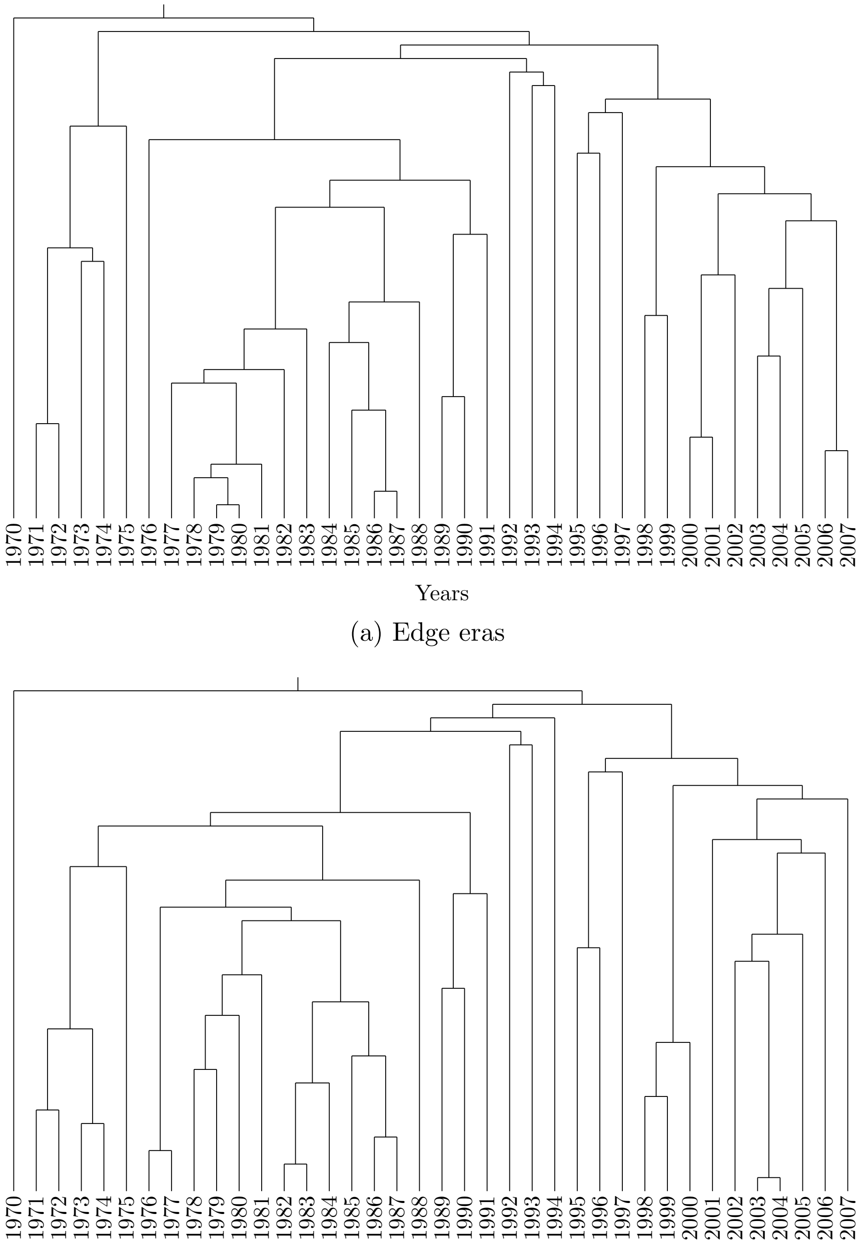


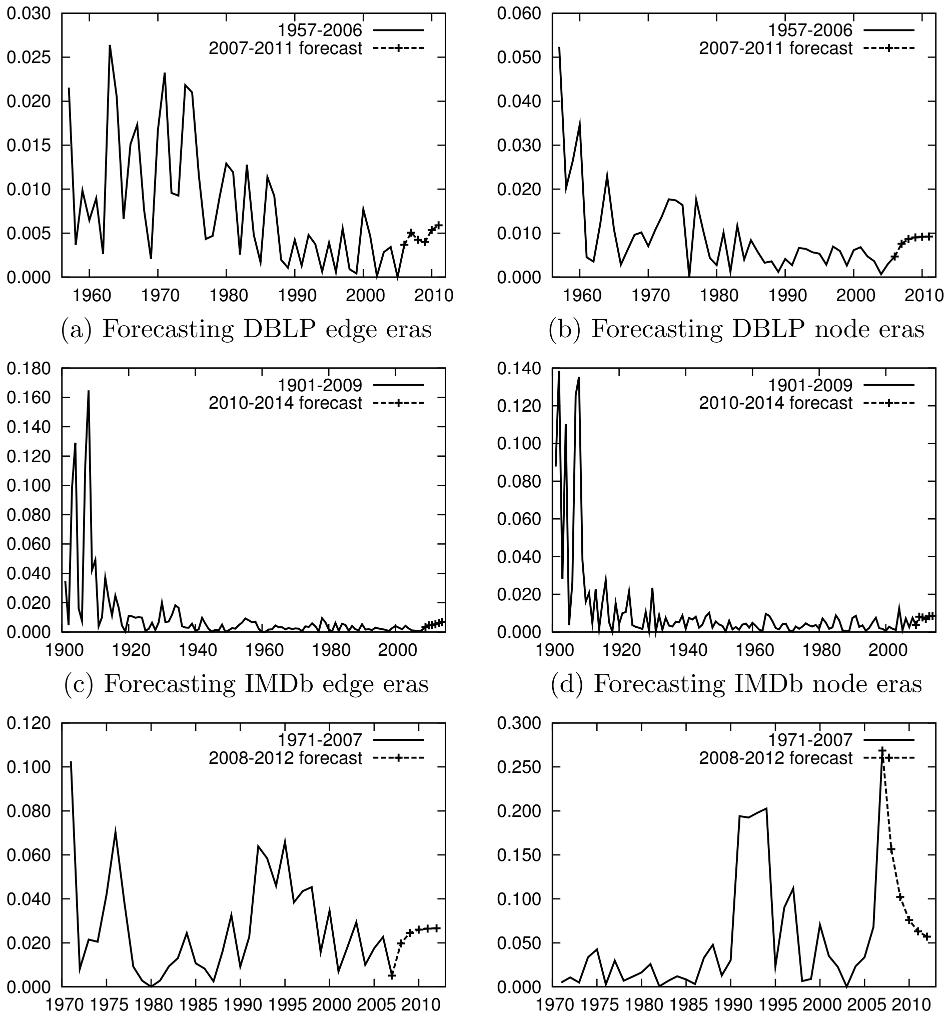
Figure 5: Eras in IMDb node evolution



(a) Edge eras

(b) Node eras

Figure 6: Eras on both edge and node evolutions in GTD



(a) Forecasting DBLP edge eras

(b) Forecasting DBLP node eras

(c) Forecasting IMDb edge eras

(d) Forecasting IMDb node eras

(e) Forecasting GTD edge eras

(f) Forecasting GTD node eras

Figure 7: Forecasting eras on dissimilarities via autoregressive models