Algorithms for analyzing evolving networks on the Dark Web & in science

Boekhout, H.D.

1 Introduction

1.1 Real-world networks

People are increasingly being confronted with large scale data collection and analysis in their day to day lives. For example, while browsing the web they are confronted with choices about which cookies to allow or whether they wish to share their geolocation and other types of data when setting up new devices or apps. The resulting wealth of large datasets, also referred to as "big data", has become the subject of many commercial applications and scientific studies. Traditional data analysis usually considers each *data record* as a separate and largely independent piece of information relating to an entity or event. Traditional methods for data analysis aim to find patterns among such records, but often overlook how these records may interact. In order to analyze the interaction between entities, the data can be modeled as a *network*, also known as a *graph*. Network modeling is a data modeling technique that tries to capture the complex interactions within real-world systems. A network model defines a set of entities within the system, which are referred to as *nodes* or *vertices*, and connects these entities based on observed interactions or relationships between them, which are referred to as *links*, *edges* or *ties*.

Data is often collected or made available at the level of a node or edge of the modeled network. For example, consider a system of roads as shown in Figure 1.1. Fastest routes to travel between destinations along the roads have to be determined practically everyday by, for example, navigation software. Based on an analysis of individual addresses or road (inter)sections, e.g., through traditional data analysis, it is not possible to determine such a route. After all, each address or road (inter)section knows at best their directly neighboring addresses or road (inter)sections. However, when modeled as a road network, where the network models the connections between

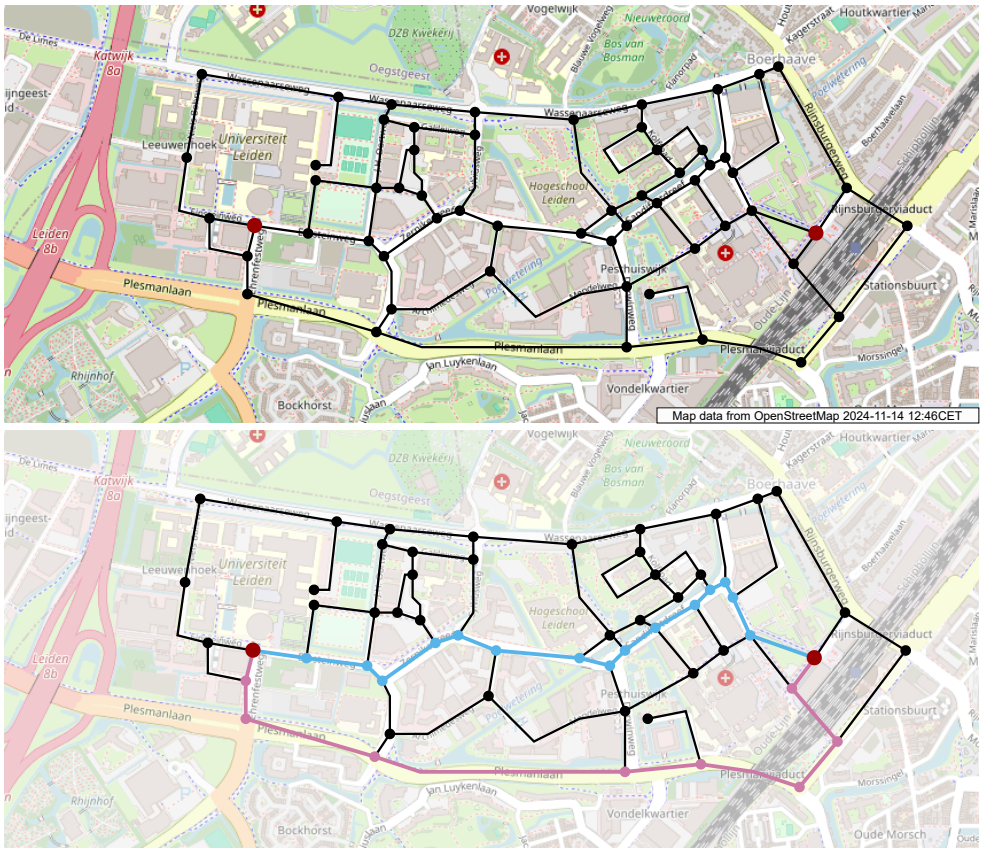


Figure 1.1: A map of the Leiden Bio Science Park as a network, with intersections as nodes and cycling paths and roads as edges. The bottom panel demonstrates how shortest path analysis can be used to infer the shortest route between, for example, the building which houses the Leiden Institute of Advanced Computer Science (LIACS) and Leiden central station. Here, the unweighted shortest path (9 edges with a cumulative distance of ≈2.1km) is shown in pink and the weighted shortest path (16 edges with a cumulative distance of ≈1.7km), with weights respresenting the true distance between intersections, is shown in blue. The map data used in this figure was obtained from OpenStreetMap (see openstreetmap.org/copyright).

addresses or road (inter)sections, it is possible to determine the fastest route. In fact, on such a network, determining a fastest route is as simple as determining the *shortest (weighted) path* between the start and end node, i.e., the sequence of connected edges that together connect these nodes with the fewest edges (or least cumulative weight). Thus, analyzing networks by, for example, looking at shortest paths, but also through

finding common interaction patterns, or by computing what the most crucial nodes and edges are, can provide insights that traditional data analysis cannot.

Systems that can be modeled as networks are everywhere. From the roads we drive on, to the social media websites we use to communicate, to our face-to-face interactions. Each can be modeled as a network. Moreover, there are numerous ways to model these real-world systems as networks, even before considering additional real-world complexities. We may consider road intersections as nodes and connect them with edges based on the roads, or instead we could consider road sections, representing a collection of addresses, as nodes connecting adjacent sections. In other words, the roads in a road network may be modeled as either the nodes or the edges of the network. On social media, if we consider users as nodes, we may connect them through following/friendship relations or connect them through mentions of one another; or instead we could consider posts as the nodes and connect them if they were viewed or posted by the same user(s). In short, there are a myriad of modeling decisions to be made when a real-world system is modeled as a network.

Modeling decisions range from decisions on what a node and edge should represent, to more complex decisions on whether to include information on the temporal order of events or the strength of connections. A common modeling decision is to include complex information but then simplify the network by means of, for example, coarse graining and aggregation. However, the specific modeling decisions that are made, are often dictated by the research question under consideration. For example, in the road network example from Figure 1.1, our research question boils down to "What is the shortest route connecting points A and B?". Since any travel must follow the roads, representing intersections as nodes and the roads connecting them as edges prevents any two edges from covering overlapping portions of the road. Points A and B can then be chosen as the closest intersection or can be represented by new nodes in-between the intersections that connect the roads they lie on. The length of each road section can then be modeled by edge weights such that any “shortest route” can be determined by weighted shortest paths. A second consideration for modeling decisions, is that the modeled network should properly reflect reality. For example, if on the map in Figure 1.1 we want to model a next-door neighbor network instead of a road network, we must first decide what constitutes a neighbor. After all, a simple choice like considering every address on a road section to be neighbors, would lead to neighbors that may live kilometers apart being considered neighbors for larger road sections. Such a definition would not properly reflect what most people would consider real next-door neighbors. We henceforth consider a network model *meaningful* when it both reflects reality properly and is fit for answering the relevant research questions.

Because a large variety of systems can be modeled as networks, the study of networks has drawn interest from various disciplines. The rapidly growing multi-

disciplinary scientific field that concerns itself with the study of and computation on networks, is known as network science [138]. Although graphs and networks have been studied for longer, the field of network science, as it is known now, was not established until the late 1990s. The field rose to prominence when scientists came to better understand the behavior of real-world networks through, for example, the introduction of the "small-world" model [151] and "preferential attachment" [10]. A key finding was the fact that the organization of real-world systems into networks is not "random", but follows a set of universal laws in terms of how connections are formed. The fact that many networks have properties such as these in common, has led to growing interest in studying network models. After all, since commonalities exist between networks representing systems from a wide variety of domains, methodologies and insights are also more widely applicable. However, network science is still a relatively young field of study with many remaining challenges. Not the least of these challenges is how to best deal with the complexities introduced by temporal data, i.e., how to analyze networks that change over time. This thesis deals exclusively with data that include some form of temporal information.

Incorporating temporal data into the study of networks has seen significant interest in recent years. In 2019, Holme and Saramäki [72] published a book mapping out the many varying approaches to temporal networks in the literature. They identified three main research themes within the study of temporal networks: (1) simplifying and coarse graining; (2) identifying important nodes; and (3) how structure affects dynamics. In this thesis, as we rely on a variety of approaches to using the temporal information in our efforts to answer relevant real-world questions, we deal with each of these three research themes. Figure 1.2 explicitly positions the chapters of this thesis with respect to these three main research directions.

The remainder of this introductory chapter is structured as follows. First, in Section 1.2, we discuss network science theory and methods. Then, Section 1.3 introduces the two real-world systems studied in this thesis and discusses the complexities of modeling these systems as networks such that the models are meaningful. Next, we discuss data quality issues that may occur for data collected on real-world systems in Section 1.4. Finally, in Section 1.5 we provide an outline of the thesis, formulating relevant research questions for each chapter.

1.2 Network science theory and metrics

A network, or graph, is defined as a set of nodes connected by a set of edges. For each node, the nodes that are directly connected to it by an edge are referred to as its *neighbors* and these neighbors collectively form its *neighborhood*. Each edge

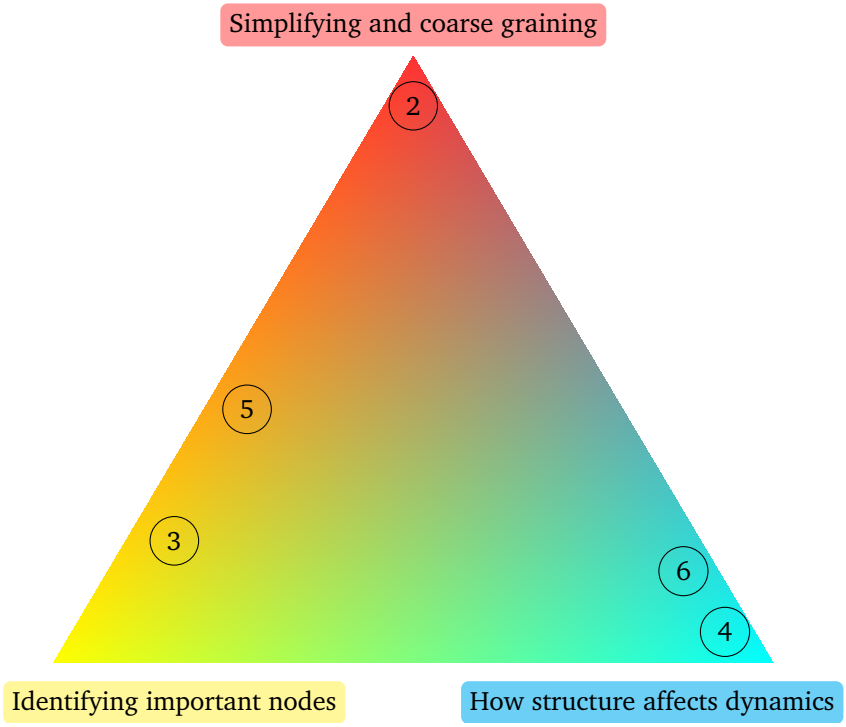


Figure 1.2: A mapping of the thesis chapters with respect to the three main research themes as described by Holme and Saramäki [72].

connects, for a simple graph, one pair of nodes. Consequently, although some nodes may not be directly connected by an edge, they may be indirectly connected through a sequence of edges. Such a sequence of edges is referred to as a *path*. Assuming a simple unweighted graph, a *shortest path* connecting two nodes is a shortest sequence of edges that connects them. The length of such a shortest path between two nodes, i.e., the number of edges that it consists of, is referred to as the *distance* between these two nodes. Shortest paths and distances are at the basis of many more advanced network metrics [11], which we will discuss in Section 1.2.2.

A *subgraph* of a graph is the network induced by a subset of nodes from the full network. This induced network may include all edges connecting the subset of nodes that are in the full network or a subset of them. *Weakly connected components* are induced subgraphs that include all edges. For weakly connected components it must hold that every node in the subset is reachable from every other node by an undirected path. The largest weakly connected component is usually referred to as the *giant component*. In many real-world networks, this giant component encompasses

almost the entire network and is often used as the subject of study in place of the full network. *Strongly connected components* differ from weakly connected components in that, instead of requiring an undirected path to exist, they require a directed path in both directions to exist between every node pair. However, finding these paths requires directional information not present in a simple graph. Therefore, we first discuss more advanced network types, such as directed networks, in Section 1.2.1 below.

1.2.1 Network types

So far, we have described only simple graphs, i.e., networks that are undirected, unweighted, and static. However, the studies presented in this thesis require us to consider the following more complex network types.

- *Directed networks* add direction to the edges. Specifically, they define for each edge a *source* and *target* node, where the edge models an interaction/relation from the source node to the target node. This may be used to model, for example, a one-way street in the road network. In directed networks there is a difference between a *directed path*, which may only follow the edges along their directions, and an *undirected path*, which may follow edges in either direction. Unless stated otherwise, a path in a directed network is assumed to be directed.

- *Weighted networks* add weights to the edges. These weights may model differences in the difficulty of traversing the edge, or alternatively they may model the strength of the connection between the entities. For example, in the road network, edge weights may model the average travel time between road (inter)-sections. Computations of shortest paths require a proper understanding of what the weights in the network model. If the weights determine the difficulty of traversing an edge, a shortest path can be determined using the cumulative edge weights. However, if weights determine the strength of a connection, we would want to traverse higher weighted edges more easily. Therefore, in such cases shortest paths would need to be computed using the inverse edge weights instead. Note that *unweighted* metrics ignore weights.

- *Temporal networks* (or *evolving networks*) add timestamps to edges (or nodes). These timestamps imply a temporal ordering of events, i.e., they indicate when the interactions/relations take place. Note that networks without timestamps are called *static networks* and that static metrics applied to temporal networks ignore the temporal information provided by timestamps. Depending on the system that is being modeled, timestamps can have different implications. They may indicate the moment that a connection is formed with the assumption it continues to exist, leading to an ever growing network; or they may indicate

the exact moment an event/interaction occurred without the presumption of perseverance of the link. Additionally, alongside the timestamp a duration may be defined to indicate how long the link exists. Thus, there are multiple types of temporal networks that each model temporal information slightly differently. A more elaborate overview is given in [72]. This work highlights how designing metrics for temporal networks is not straightforward, and that newly designed metrics may only be applicable to one type.

One common approach to simplifying a temporal network such that existing static metrics may be applied, is to create a *snapshot network*. A snapshot network divides the entire lifetime of the network, i.e., the range of timestamps, into time intervals. If edges are presumed to persevere, each snapshot aggregates the temporal network to a static network using all edges up till the end of its time interval. Thus, including those preceding the time interval. If edges do not persevere, each snapshot aggregates the temporal network only for those edges with a timestamp within its interval. By applying static metrics to the various snapshots, it is then possible to study how the network evolved over time. Of course, the individual snapshots do discard knowledge of the exact order of events, leading to some information loss.

Different approaches to temporal networks serve different purposes. This is reflected in our use of temporal networks in this thesis. In Chapters 4 and 6 we study a temporal network directly. On the other hand, in Chapters 3 and 5 we create snapshots, the former with persevering edges and the latter without.

- *Multiplex networks* are a limited form of *multilayer networks* [81]. Both allow for the possibility of multiple types of edges, each modeling a different type of interaction/relationship. However, where multilayer networks also allow for each layer (i.e., edge type) to have their own separate set of nodes, multiplex networks are defined to use the same node set for each edge type. In this thesis we do not compute directly on any multiplex networks, but do simplify a multiplex network to a single layer (uniplex) network in Chapters 2 and 3.

All network types described above may of course be combined into one aggregate type. Moreover, it should be noted that more network types have recently been distinguished, including simplicial complexes or higher-order networks [14, 15, 16].

1.2.2 Network metrics

Networks metrics summarize particular structural features of the network, and can be divided into three categories of scale: micro-, meso-, and macro-level measures [111, 59]. Micro-level metrics are at the level of nodes and their direct neighborhood. For

example, the *degree* of a node indicates the number of neighbors of a node. In a directed setting, a distinction can be made between the *in-degree* and *out-degree* of a node, indicating respectively the number of neighbors that have a link to the node and the number of neighbors the node links to. In a weighted setting the degree may instead refer to the cumulative weight of connections to the neighbors, then commonly referred to as *weighted degree*, while a multilayer setting may define degrees for each layer separately. Many micro-level features can also be directly obtained from the data records underlying the network.

Measures that consider the entire network, are considered macro-level. These measures describe the network as a whole. Examples of such metrics used in this thesis are: the *degree distribution*, which determines the frequencies of the node degrees; the *average degree*, which computes the average node degree; the *density*, which measures the proportion of possible edges that actually exist; the *average clustering coefficient*, which measures the extent to which nodes cluster together in the form of triangles; and the *diameter*, which measures a longest shortest path between any two nodes in the network. Macro-level metrics often serve as a means of understanding its general structure, to categorize networks, and to improve understanding of how the network structure may be affecting algorithm performance.

The meso-level includes network measures and structures that consider more than the direct neighborhood but not the entire network. Examples of such structures are subgraphs such as weakly and strongly connected components of a network. Other examples of network subgraph structures that are considered meso-level are *cliques* and *motifs*. Cliques are subsets of nodes that are fully connected, i.e., every node in the subset is connected to every other node. On the other hand, motifs are subsets of nodes that are not necessarily required to be fully connected, but are (usually) required to be at least weakly connected. The term motifs is used to denote frequently occurring subgraph structures and often specifically those that occur more frequently than expected. The study of cliques and motifs provides insight into common interaction patterns and group-dynamics within complex systems [27]. Finally, another well-known and often studied meso-level structure, is the network *community*. Communities are subsets of nodes that are more densely connected within the communities than they are connected to nodes outside the community.

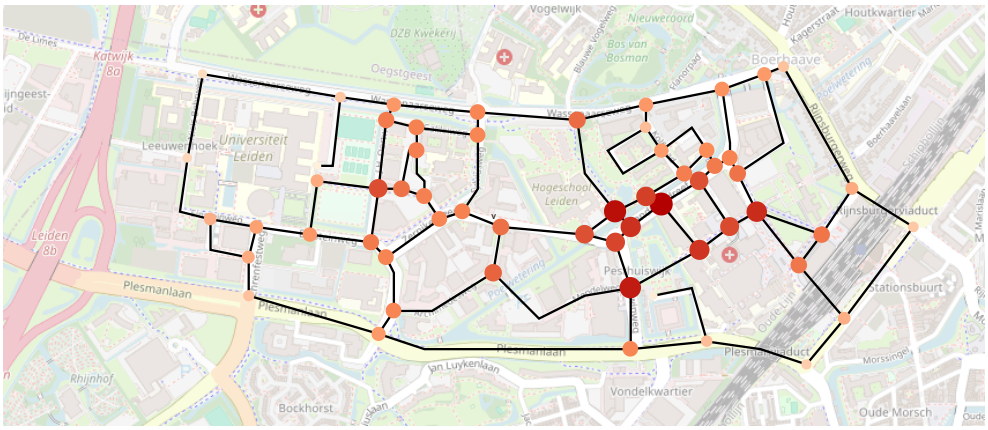
Centrality measures, metrics that measure the positioning of a node within the network, combine meso- and macro-level information into micro-level metrics. In this thesis we employ the following centrality measures.

- *Degree centrality*, which simply ranks nodes based on their degree.
- *Eigenvector centrality* [125], which derives the importance (i.e., centrality) of a node not just from how many neighbors it has, but also from how important

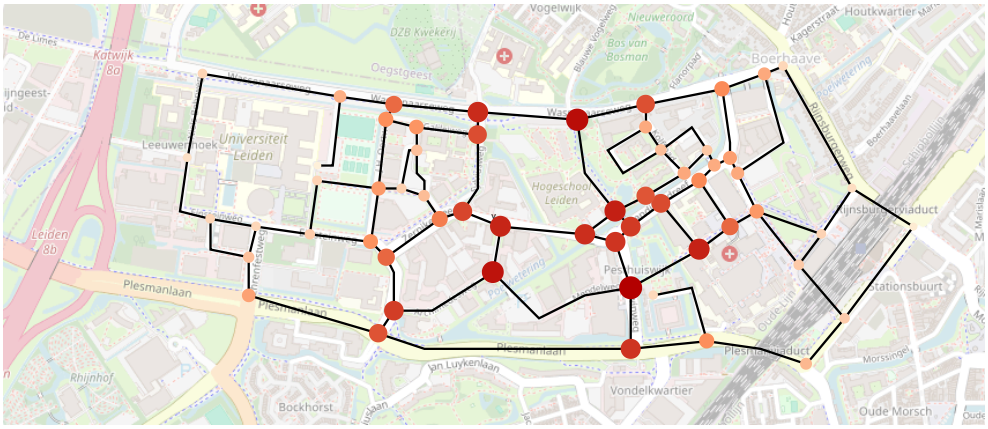
those neighbors themselves are.

- *Closeness centrality* [124], which measures how easily a node can reach all other nodes in the network. Essentially, it computes the inverse of the average shortest distances to all other nodes.

- *Betweenness centrality* [33, 68], which measures the extent to which a node connects the network. Specifically, it computes how often a node lies on a shortest path connecting any two nodes in the network.



(a) Eigenvector centrality

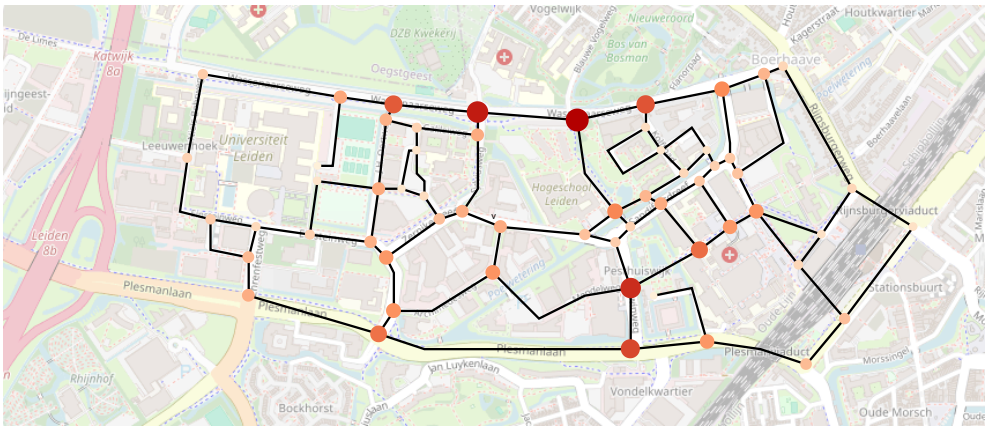


(b) Closeness centrality

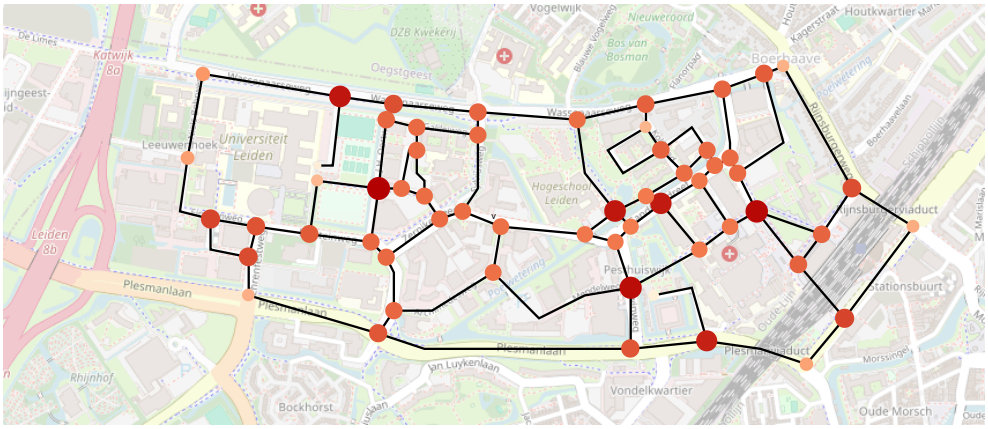
Figure 1.3: Eigenvector and Closeness centrality measures applied to the unweighted road network from our example in Figure 1.1. The larger the node size and the darker the node colour, the more central the node.

- *PageRank centrality* [114], which computes the probability that a random walker ends up at a given node. This random walker always either follows an edge to a neighboring node or, with a particular probability, jumps to a random node instead. The possibility of jumping to a random node is the main difference between PageRank and eigenvector centrality.

Figures 1.3 and 1.4 give a visual representation of each of these centrality measures, highlighting the most central nodes for each.



(a) Betweenness centrality



(b) Pagerank centrality

Figure 1.4: Betweenness and PageRank centrality measures applied to the unweighted road network from our example in Figure 1.1. The larger the node size and the darker the node colour, the more central the node.

1.3 Introduction to two real-world networked systems

In this thesis, we study two real-world systems more closely: *dark web cryptomarkets* and the *scientific research and publication system*. Here, we provide an introduction to both systems and discuss what issues arise when modeling these systems as networks.

1.3.1 Dark web cryptomarkets

Dark Web Markets (*DWMs*), also called *Dark Net Markets*, are online marketplaces on the dark web which often facilitate trade of illegal goods and services. Such illegal goods and services, can range from drugs and personal information (e.g., ID, credit-card, etc.) to weapons and assassinations to illegal pornographic material. DWMs are becoming increasingly popular due to the anonymity they provide their users. This anonymity is facilitated by allowing access to the DWM only through browsers such as *The Onion Router* (*TOR*) [58] and by facilitating payments with crypto currencies such as Bitcoin [110]. We refer to a DWM as a *cryptomarket* when they provide a moderated forum for their users, and escrow services. These additional services can improve the trust between vendors and their customers by reducing the likelihood of scams, often leading to increased trade. Given the largely illegal nature of the goods and services traded on cryptomarkets, law enforcement agencies want to better understand and disrupt these markets [128].

Figures 1.5 and 1.6 show example reconstructed screenshots of, respectively, a listing and a forum page of a cryptomarket. As shown in Figure 1.6, the forum of a cryptomarket is often used by vendors to promote their listings and by customers to review those listings. Therefore, the communication on the forum is often indicative of market trade. Research into this forum communication has primarily focused on text/sentiment analysis [108, 137, 105, 2, 79, 95]. However, the criminal world increasingly uses encrypted communication. Moreover, interactions within online ecosystems are known to be informative for understanding its users [152]. Additionally, text and network analysis have been shown to provide complementary insights when combined [12]. Anticipating the needs of law enforcement agencies for methodologies in a context where textual information is not readily at hand, this thesis investigates whether network analysis of communication on cryptomarket forums may be of use to law enforcement activities.

To model the communication on a cryptomarket forum we can create so-called *communication networks* that capture who communicates with whom in a system. On a forum, we can consider two users to be communicating when they post within the same topic. Poquet et al. [117] showed that, while user posting behavior is highly predictive of their degree in forum communication networks and may therefore not reflect social

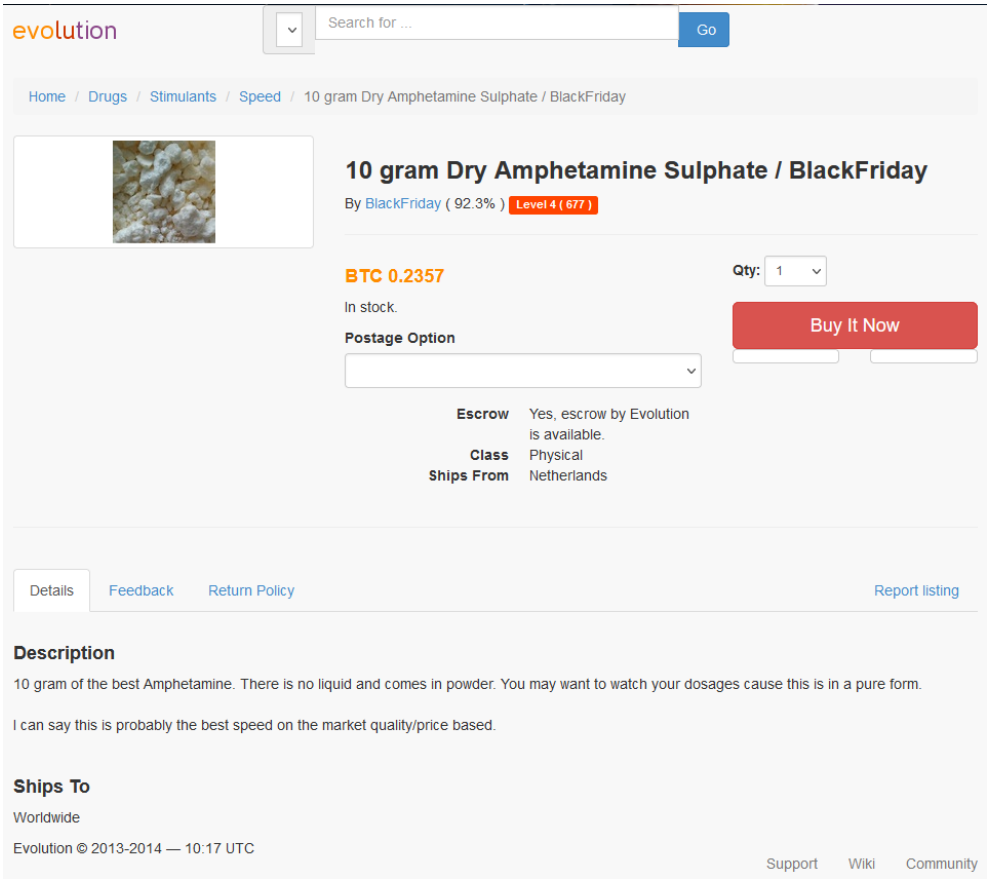


Figure 1.5: Screenshot of a listing page on the cryptomarket EVOLUTION.

dynamics, the structure of the network captured through, for example, the weighted local clustering coefficient, is a better approximation of social relationships. In other words, meso- and macro-level metrics as discussed in Section 1.2.2, may provide new insights not captured through posting behavior for forum communication networks.

A straightforward modeling approach for a forum network would be to simply connect everyone posting in a topic. However, such an approach would not match reality. After all, there is nothing requiring a user to read all previous posts in a topic. Additionally, there is nothing requiring a user who previously posted to read newer posts. Thus, the question becomes, how much time may have passed and how many posts may have been placed in between, for us to consider two posts in the same topic to indicate likely communication between users? Furthermore, should we consider a quick reply a stronger connection or not? We may also wonder, does every post

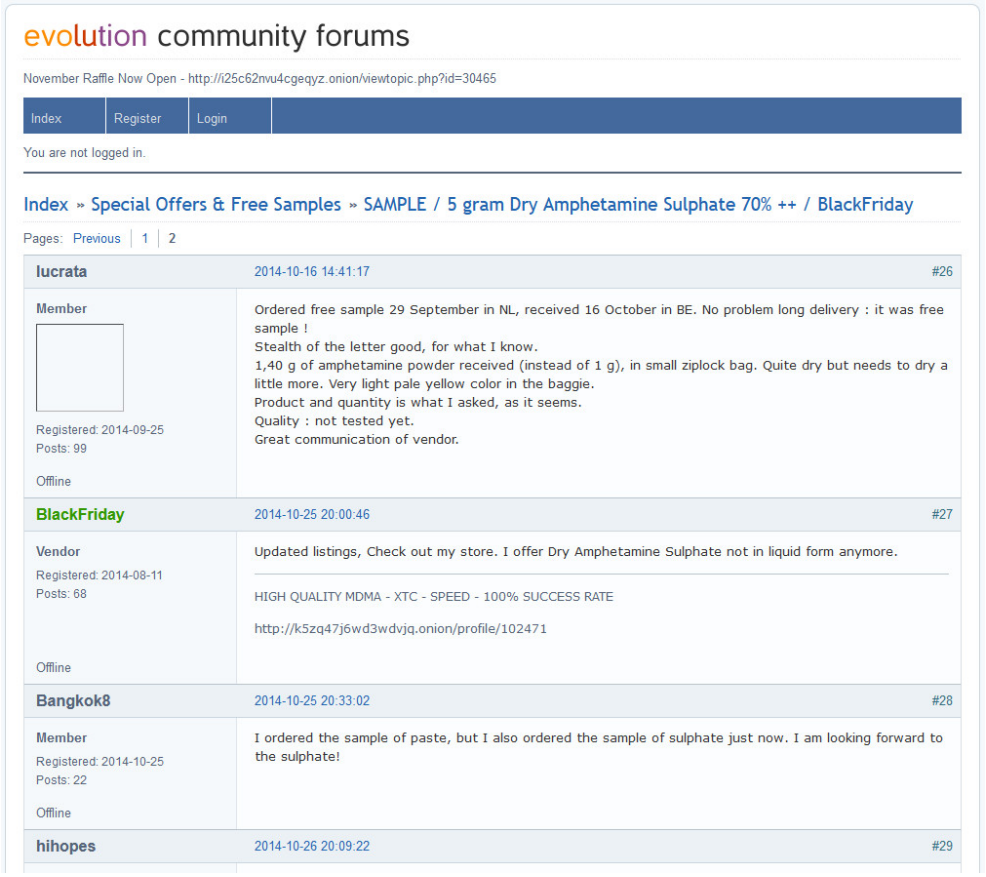


Figure 1.6: Screenshot of a forum page of a topic promoting a sample listing on the cryptomarket EVOLUTION.

constitute a form of communication to whomever started the topic? After all, if a vendor started a post to promote a listing, it would be reasonable to assume the posts in the topic are aimed at or about that vendor. Finally, we need to consider the temporal ordering of events. Since each communication link is formed based on two temporal events (i.e., two posts), which should be considered the timestamp? Can we reliably and meaningfully form time-respecting paths, i.e., paths where each subsequent edge must have a higher timestamp, from such temporal information? Or should we stick to a static analysis? Each of these questions addresses a modeling decision that is directly relevant to the interpretability and meaningfulness of results. We tackle each of these (and more) questions in Chapters 2 and 3.

1.3.2 Scientific research and publication system

The *scientific research system* is made up of researchers at universities and research institutions that alone or in collaborative teams perform research to answer scientific questions. The primary outputs of the scientific research system are publications, in which the researchers describe their research and present their results and conclusions. In the past century, the research system has seen an exponential growth in the annual number of publications [67]. Due to such rapid growth, there is a need for governments, universities, and institutions to better understand the research system and how it has evolved, in order to make appropriate policy decisions. Consequently, the interdisciplinary research field sometimes referred to as the Science of Science [67, 146], explores the underlying mechanisms of and relationship between scientific impact [145, 135], productivity, collaboration [23, 135], career trajectories [17], and the fundamental workings of science.

Research into the scientific system consists of both *qualitative* and *quantitative* research. Quantitative research focuses on data and numerical analysis to test hypotheses and detect patterns, whereas qualitative research explores experiences and meaning through methods such as interviews. Quantitative research into the scientific system most often deals with publication data. Over the years, several large-scale publication data repositories, both commercial and open access, have emerged. Such publication data repositories are also known as *bibliometric databases*. Examples of such bibliometric databases, that are well-known and often used in research, are Clarivate's *Web of Science* (*WoS*), Elsevier's *Scopus*, and the *OpenAlex* databases [118].

Publication records often include at least the *authors* and their *affiliation(s)*, i.e., the university or institution where they were employed while performing the research presented in the publication. Many other pieces of relevant information may be included, such as: keywords; abstract; full text; affiliation address; author contribution statement; funding information; citations/references; etc. In this thesis, we focus on using the author, affiliation (address), and citation information.

The two most common approaches to modeling the scientific research system as a network are to create so-called *co-authorship networks* and *citation networks*. Co-authorship networks model scientific authors as nodes, connecting them by an edge when they co-authored a publication. Alternatively, nodes can be defined at the level of organizations, scientific disciplines (i.e., fields), or countries. In such cases, the edges are aggregated, i.e., the edges connecting authors that are associated with two different nodes are combined into a single (weighted) edge. As previously stated in Section 1.1, the chosen level of aggregation most often depends on the specific research question under consideration. Citation networks instead model publications themselves as nodes, which are then connected if one publication references/cites the

other. In this thesis, we only look at co-authorship networks, but do so at two levels of aggregation: the city level and the author level.

For city level aggregation, it should be noted that there is no consensus in literature on how the spatial unit "city" should be defined [50]. Relying directly on city names from address data could easily lead to unfair comparisons when comparing between smaller cities like Leiden and megacities like New York. Urban agglomerations are seen as a solution to this issue by creating a more universal comparison criteria [50, 100]. In Chapter 5 we create a new spatial unit, i.e., urban agglomeration, for use as an aggregation level that best fits our research question.

When constructing a co-authorship network, in addition to deciding at which aggregation level we want to model, we must decide what additional information we want to include in the model. For example, do we want to add a weight representing the number of co-authorships? Or should the weight represent their affiliation distance instead? Or should weights be used to represent some other attribute of co-authorships? With respect to the temporal domain, we must decide whether we want to aggregate over all time, or only within respective publication years preserving the temporal information. Then, if we preserve the temporal information, we still need to decide whether to create snapshots and whether edges should persevere to subsequent snapshots. Unlike for the cryptomarkets, however, we do not need to account for any uncertainty about the existence of a co-authorship edge. Each of these choices are closely tied to the research question at hand and there is therefore not one correct approach. Scientific co-authorship networks and the questions posed above are further studied in Chapter 5 and 6.

1.4 Data quality

In the previous section we introduced two real-world systems from which we use data in this thesis and discussed how they could be modeled as a network. However, real-world data often comes with (possibly severe) data quality issues [45]. Data quality issues may directly impact the extent to which we are able to model the full system, due to a lack of complete information, or may affect our available modeling decisions, due to issues with inconsistent, inaccurate, or out-of-date data. The following are the five central data quality issues as described by Fan & Geerts [65].

- *Data inconsistency* refers to inconsistencies in the data records that are associated with the same entity or event. For example, cryptomarket records indicating the same user id may indicate the username "Bob" in one record and the username "Dave" in another. Data inconsistencies may therefore refer to actual changes occurring in the system or may indicate an error during data collection.

- *Duplicate data* refers to data records that are associated with the same entity or event. Especially when data on a system is collected at different moments in the lifetime or collection period of the system, does duplicate data show up often. Duplicate records may contain only identical information, they may contain complementary information that should be combined, or they may contain contradictory information that needs to be resolved. The process of resolving duplicate data issues is referred to as *data de-duplication*.

- *Data accuracy* refers to whether the values recorded are accurate with respect to the real values of the entities and/or events of the system; and if not, their "closeness" to the real values. Inaccurate data is near impossible to resolve unless there is also a data inconsistency that includes (more) accurate information.

- *Incomplete information* refers to missing data records and missing information within data records. For example, bibliometric databases are unlikely to capture every article that is published. The Web of Science database, for example, is well-known for having relatively poor coverage of conference proceedings. Yet, if one wants to study the field of computer science, in which this is a publication format as common as journal publications, this becomes problematic. Some incomplete information can be resolved due to otherwise duplicate data, but in most cases it is not possible to recover this information. When analyzing real-world systems, it is important to be aware of what data is missing and how that may affect the results and conclusions of a study.

- *Data currency* refers to the accuracy (or *timeliness*) of the data records due to uncertainty or errors in the temporal information (i.e., the timestamp). For example, errors in detection of when an event occurs may directly affect how the events are ordered and subsequently interpreted. Additionally, when information on an entity is not captured at every timestamp, uncertainty is introduced concerning how accurate and up-to-date (i.e., current) entity information is.

Naturally, the above data quality issues affect the datasets we study in this thesis as well. For the scientific research system, the bibliometric databases already resolve many quality issues. In particular the database that we use in this thesis, has been heavily curated by the Centre for Science and Technology Studies (CWTS) to resolve important data quality issues such as author name ambiguity [37, 143]. However, the data we retrieved on the dark web cryptomarkets had not previously been processed to resolve its many data quality issues. Chapter 2 explains how we resolved many of them and investigates the completeness of our dataset.

1.5 Thesis outline

This thesis consists of three parts. Part I deals with the study of dark web cryptomarkets (see Section 1.3.1). Specifically, it deals with the cryptomarket called EVOLUTION. EVOLUTION was a cryptomarket active from January 2014 until March 2015 when it closed due to an exit-scam [56]. Like all cryptomarkets, EVOLUTION consisted of a combination of a marketplace and a forum. Over the lifetime of the cryptomarket, researchers collected data from EVOLUTION (and other DWMs) by scraping both the marketplace and forum, i.e., by collecting the source files used to generate the web pages. In 2015, this data was published online as the *Darknet Market Archives* by Branwen et al. [34].

In Chapter 2, we process the raw source files of the EVOLUTION cryptomarket, that were published in the Darknet Market Archives, into a longitudinal structured dataset and communication network. In doing so, this chapter answers questions such as:

- What data quality issues exist in the forum and market data that can be extracted from the raw EVOLUTION source files? And how can we resolve these issues?
- Can we link the EVOLUTION marketplace data to the EVOLUTION forum data? How complete is the resulting dataset?
- How can we extract a forum communication network from this data, such that the edges capture interaction in a meaningful way?

In Chapter 3, we subsequently use this dataset to develop forum activity and network based methods that may aid law enforcement agencies in focusing their efforts on the cryptomarket users that are the most likely to become successful vendors in terms of sales. Furthermore, we study the cryptomarket at different points in time to determine whether the current state of the forum can be indicative of future vendor success, such that law enforcement may employ preventative measures instead. We answer the following questions:

- When ranking users based on forum activity and network centrality measures, which metric detects the most current (and future) successful vendors among highly ranked users?
- Are network centrality measures able to find successful vendors that forum activity measures can not?
- Are either forum activity or network centrality measures capable of serving as an early warning signal for future vendor success?

Part II (Chapter 4) provides a fundamental contribution to network science by introducing a new fast temporal maximal clique enumeration algorithm. Current state-of-the-art methods enumerate so-called (δ, γ)-maximal cliques, where the cliques must be fully connected for every δ-time period in its time span with a minimum frequency of γ for every node pair. For our algorithm we introduce a new temporal clique definition with two goals: (1) to resolve a problem in the temporal domain produced by the existing definition; and (2) to extend the application from temporal networks to weighted temporal networks (see Section 1.2.1). Consequently, the latter goal allows for the inclusion of meaningful information about the strength of connections, or difficulty of edge traversal, instead of only the frequency of these connections. Our new algorithm can be applied on larger networks, with millions of edges, with a reasonable runtime, i.e., hours instead of days, provided that the network is not too dense. Consequently, this improvement allows us to apply temporal maximal clique enumeration on the global persistent co-authorship network, with over 70 million temporal edges, in Chapter 6. Chapter 4 answers the following questions:

- Can we efficiently enumerate (δ, γ)-maximal cliques where γ refers not to the frequency but the cumulative weight of connections, i.e., cliques that are fully connected for every δ-time period in its time span with a minimum cumulative weight of γ and such that the cliques are both temporally and size-wise maximal?
- Can we improve the overall speed, i.e., efficiency, by which (δ, γ)-maximal cliques are enumerated, while still guaranteeing the completeness of enumeration?
- How efficient is our proposed algorithm compared to state-of-the-art temporal maximal clique enumeration algorithms on real-world (weighted) temporal networks as well as synthetic temporal networks of increasing structural, i.e., static, and temporal density?

Part III deals with publication data from the scientific research system (see Section 1.3.2). In Chapter 5 we study the evolution of scientific co-authorship networks that are formed at the level of cities. In order to do so, we first introduce a new spatial unit we call the *scientific city*, which can be determined using affiliation address data. Scientific cities are groups of addresses that are physically close. As a spatial unit, the scientific city therefore lies between the organization/affiliation and the named/normal city level, such that it represents a spatial distance where collaboration remains easy due to short travel times. Through an analysis of the most central scientific cities in the networks over time, we answer the following questions:

- How can we determine a spatial unit, fit for a fair comparison of the scientific system, that is larger than the organization/affiliation but that (largely) remains within given city limits?

- How much change over time is there in the most important (i.e., central) scientific cities?

- Can we quantify the stability of the network by means of a comparison to rewired networks? And what does this imply in terms of where in the network new connections appear (or disappear)?

In Chapter 6 we investigate the citation success of scientific teams. Teams have been shown to produce more highly cited and high-impact research than scholars working independently [155]. However, an apparent paradox has emerged in the literature looking at the success of teams. On the one hand, persistent collaboration was shown to be beneficial to a team's success [132]. On the other hand, fresh teams with few pervious collaborations were shown to improve success [156, 96]. We investigate the relationship between team freshness, persistence, and their success by looking at the success of *persistent scientific teams*, i.e., sets of scholars that co-author sufficiently many publications within a sufficiently short time span. Persistent scientific teams are thus formed by temporal maximal cliques of persistent collaborating scholars. To determine these teams, we relied on the maximal clique enumeration algorithm presented in Chapter 4. By investigating the composition of persistent scientific teams and when their success occurs, as well as the temporal relationship with teams with overlapping members, we answer the following questions:

- How can we determine scientific teams that capture the fluid and interdisciplinary nature of modern science?

- Is it common for scientific teams, that collaborate and publish together often, to appear among the author teams of scientific publications, i.e., are persistent scientific teams prevalent?

- Are successful persistent scientific teams more likely to produce highly cited works early or later in their collaboration?

- Are authors of successful persistent scientific teams more likely to be more institutionally and geographically diverse?

- Do related persistent teams, i.e., teams with overlapping members and temporal overlap, impact a team's success? For example, do new team formations help keep a team fresh? And does prior collaborative experience aid the new collaboration?

Finally, we conclude the thesis in Chapter 7.

Publications

The chapters of this thesis are based on the following manuscripts and peer-reviewed articles:

Chapter 2 H. D. Boekhout, A. A. Blokland, and F. W. Takes. A large-scale longitudinal structured dataset of the dark web cryptomarket Evolution (2014–2015). *arXiv preprint arXiv:2311.11878*, 2023

Chapter 3 H. D. Boekhout, A. A. Blokland, and F. W. Takes. Early warning signals for predicting cryptomarket vendor success using dark net forum networks. *Scientific Reports*, 14(1):16336, 2024

Chapter 4 H. D. Boekhout and F. W. Takes. Fast maximal clique enumeration in weighted temporal networks. *Social Network Analysis and Mining*, 16(1):10, 2026

Chapter 5 H. D. Boekhout, E. M. Heemskerk, and F. W. Takes. Evolution of the world stage of global science from a scientific city network perspective. In *Complex Networks & Their Applications X*, pages 142–154. Springer International Publishing, 2022

Chapter 6 H. D. Boekhout, E. M. Heemskerk, N. Pisani, and F. W. Takes. Freshness, persistence and success of scientific teams. *arXiv preprint arXiv:2507.12255*, 2025 (under review)

A full list of publications by the author, including those that did not form the basis of chapters in this thesis, can be found on page 229 of this thesis. Data and code availability statements have been incorporated within the chapter texts and can be found in Sections 2.3, 2.5.2, 3.2.1, 4.5, 4.6.1, 5.3.1, and 6.4.1. Additionally, acknowledgments of our primary bibliometric data source (CWTS) have been included in both the front- and back-matter of this thesis.