Improved e-mail forensic using dynamic graphs and change-point detection

Christian Hiller and Andreas Wagner

Hochschule Karlsruhe - Technik und Wirtschaft
andreas.wagner@h-ka.de

Abstract. Fraudulent behavior costs the German healthcare system an estimated 14 billion euros per year. Reasons are, amongst others, criminal networks of nursing services, doctors and patients. To investigate such cases, authorities often examine the e-mail communications of suspects. This still requires very high effort in practice, as often all e-mail communication is actually read manually. This work proposes algorithms based on graph metrics and change-point-detection to automatically identify changes in the communication structure of e-mail accounts over time. This can speed up investigations, as it enables authorities to reduce the amount of data to evaluate manually. The starting point for the proposed method is a dynamic graph modeling of e-mail communication. Then graph metrics are calculated and the resulting time-series of graph metrics are analysed using change-point detection methods. An evaluation of the methods on the infamous ENRON data set shows the potential to support forensic investigations.

Keywords: fraud investigations; healthcare; e-mail forensic; dynamic graph metrics, density; average clusterin; change-point-detection.

1 Introduction

Billing fraud and corruption in the health care sector causes estimated costs of about 14 billion euros per year for the German social system [1]. In this context, corruption is often organized in network structures. Only recently, a German newspaper reported three cases of large-scale billing fraud by networks of doctors, nursing services and patients in the German cities of Munich and Augsburg. Each case resulted in damages between two and three million Euros [2]. When working on such cases, investigating authorities need to analyze large amounts of communications data in order to understand the structure of the network. They are interested in the actors involved, need to reconstruct the hierarchy of the network, try to identify important events, and, most importantly, gather legally tenable evidence.

In practice, mainly e-mail communication data is available to the investigators. In a typical fraud case there can be as many as 400.000 e-mails, which need to be taken into account. From our experience, the evaluation is often done manually, which means that a police officer is actually reading the e-mail content in order to determine valuable information. The analysis includes the evaluation of text, the temporal classification of messages, and the identification of groups, communities, and patterns in communication behavior. These activities require high capacities and lead to long investigation times. In the publicly funded research project *Kriminelle Netzwerke* [3] the authors work with German police authorities in order to develop tools and methods from mathematical graph theory to support the investigations with algorithms for data evaluation.

In this paper we introduce an approach based on dynamic graph interpretation and change-point analysis as a decision support system for investigators. It is based on the intuitive idea that suspicious elements of a criminal case are reflected in the communication behavior of participants and thus in the dynamic graph model. Automated identification of such changes can narrow down the time period and thus the amount of e-mails for investigators to examine. This frees up investigators' capacities. We propose to interpret the e-mail communication as a dynamic graph. Based on this dynamic graph representation we select graph metrics for particular use-cases of the investigator and automatically detect change-points in the structure of the network. This helps to identify conspicuous periods of time. We evaluate our approach on the famous ENRON e-mail corpus [4].

The paper is structured as follows. In Section 2 we give a summary of research on the application of graph-theory in police investigations. Section 3 introduces our approach using graph metrics and change-point detection (CPD). In a case study in Section 4 we validate the proposed approach on ENRON-data. We conclude in Section 5 with a summary and an outlook.

2 Related Work

In the context of e-mail forensics, (static) graph representations, as shown in Figure 1, have been used to analyze communication structures in previous research. There is also previous research on temporal changes in e-mail

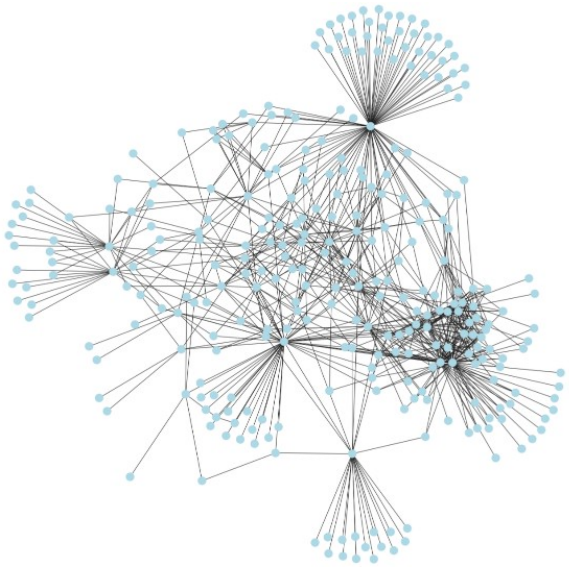


Fig. 1: Exemplary illustration of a communication network as a graph

users' communication behavior. In the following, we summarize the current state of research on those topics. The e-mail mining toolkit developed by [5], [6], and [7] is a modular framework for analyzing e-mail accounts through statistical analysis and graph representation of e-mail communications. [8], [9] and [10] present similar frameworks. They add functionalities for the classification and clustering of e-mails based on their content. The analysis of such graph models by computing graph metrics has been done in [11]. In [12] an application for dynamic graph visualization of e-mail data is presented. Although a temporal distribution of e-mails is also visualized here, a dynamic graph model is not used.

In general, *dynamic* graph models are rarely used and CPD is not applied in any of the previously cited research. In contrast, in research on social network analysis, which we evaluate in the following, dynamic graph representations and CPD are already used to identify conspicuous periods. This research concludes that structural changes in the dynamic graph representation of a social network can be indicators of changes in the real network. [13] model the e-mail communications of students and employees at a university and discuss the change in average node degree, clustering coefficient, length of the shortest path, and size of the largest connected component. [14] model the Bitcoin block-chain transactions as a graph and analyze the change of the graph over time on its adjacency matrix through principal component analysis and eigenvalue decomposition. [15] model communication through short messages from employees of a hedge fund as a dynamic graph and discuss the effects of internal and external events, such as drops in stock prices, on the structure of the graph and employee communication. It is concluded that negative events make communication more intense and the graph more densely connected. All approaches have in common a modeling of the dynamic graph in terms of discrete, successive static graphs. Studies show that real events are reflected in the structure of dynamic graph representations of social networks and that these can be captured by analyzing the time course of graph metrics. CPD approaches can identify such changes in an automated manner [16]. Advances in the fields of social network analysis and graph representation learning are continually yielding new approaches for identifying changes in the structure of graphs. The following are examples of some methods for CPD on graphs. [16] explore the identification of change points by analyzing the distribution function of the node degree of a dynamic graph. Change points are identified by computing a distance metric between the distributions of successive graphs. A hypothesis test is then used to determine the probability that each graph is a change point. [17] propose CPD on generative graph models. In addition, there are approaches to CPD on changes in the Laplacian spectrum of a dynamic graph [18] and an approach by clustering snapshots [19].

A study by [20] compares the results of CPD algorithms on traditional graph metrics with complex metrics derived from generative graph models. The authors conclude that analysis of traditional graph metrics yields only slightly worse results with significantly less complexity and computational power required. The most precise change points can be identified by the number of active nodes and edges. In contrast, the density, the average clustering coefficient and the average shortest path are robust to structural changes in a graph.

This literature review shows that graph-based methods for representing e-mail data are used in e-mail forensics. However, such frameworks rarely use dynamic graph models. Even further, work in the area of social network analysis shows that CPD approaches can be used to deduce real-world events from dynamic graph models. The

goal of this work is to transfer such an approach to the field of e-mail forensics in order to support the work of police investigations.

3 Method

To identify periods or points in time of particular interest for investigation, an approach using CPD on a time series of a graph metric is designed. This approach requires the development of a model for representing the communication structure of e-mail networks, with special consideration of the evolution of these networks over time. Based on such a dynamic model, metrics and their changes over time are computed. These time-series then serve as an input for CPD algorithms.

(1) Preparation of data
To model e-mail data as a dynamic, structural communication graph, e-mail metadata is required in a format similar to Table 1. During preprocessing, e-mails with multiple recipients are split into multiple e-mails with one recipient. The information to how many recipients the e-mail was sent is stored in the 'No. of receivers' feature.

Table 1: Features used to describe e-mail communication.

Timestamp	Sender	Receiver	No. of receivers
1997-07-10 14:05:57	student1@h-ka.de	student2@h-ka.de	2

(2) Modelling a dynamic graph
To represent the network structure of e-mail data, it is modeled as an undirected, weighted graph. Nodes in the graph correspond to e-mail addresses. An edge exists between two nodes if there is at least one e-mail exchange between the e-mail addresses. The weight of an edge corresponds to the number of e-mails between the addresses. We do not use directed graphs, which would also represent the sender / receiver structure. This is a possible extension of our work.

To account for the number of recipients of an e-mail, an e-mail contributes with $1/$*total number of receivers* to the weight of an edge. To analyze the change over time, the graph is modeled dynamically. The dynamic is represented using a series of static graphs. An element of the series, i. e. a single static graph, is referred to as a *snapshot* in the following. The snapshots are build from the e-mail corpus such that the e-mails are divided into temporally consecutive blocks of equal length. The blocks not necessarily need to be disjoint. Changes between these snapshots represents changes in the communication structure over time. Note that we keep the set of nodes the same across all snapshots. It includes all e-mail addresses in the data-set. Nodes with no edges in a snapshot are called inactive.

The snapshot approach is chosen because it has been proven effective in the reviewed work on CPD on social network graphs. The time period from the first to the last e-mail is discretized and modeled as a sorted set T, where for each $t \in T$ the e-mails of the corresponding time period are aggregated. The length of the time interval Δt covered by a snapshot is a hyper-parameter. Short time intervals allow for precise detection of changes, whereas long-term structural changes and features may not be detected due to the smaller graphs that would result. One way to circumvent this issue is to consider *sliding* snapshots. In this case, a snapshot is taken for each t, although a snapshot still summarizes a period Δt. One possible modeling approach, illustrated in Figure 2, may be to discretize the observation period to day-level and choose $\Delta t = 7$. Accordingly, one snapshot covers one week. This approach has a smoothing effect on metrics over time that are analyzed on the graph. It remains to be noted that each e-mail is contained in Δt snapshots.

(3) Calculation of graph metrics
The next step is to determine graph metrics on the snapshots. The metrics along with their meaning shown in Table 2 are proposed.

The metric is calculated on each snapshot. These values form a time series that describes the change in the graph structure over time. The choice of the metric determines which information about the communication structure is represented.

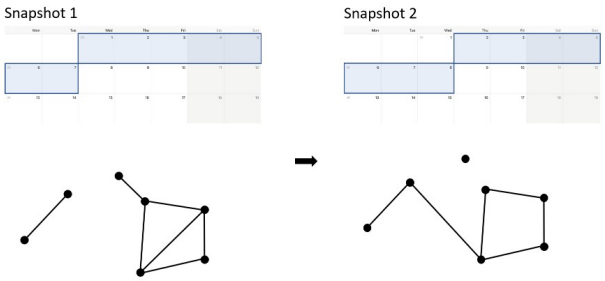


Fig. 2: Dynamic Graph Model

Table 2: Proposed graph metrics and interpretations for snapshot analysis.

Graph metric	Interpretation
Density	How closely connected are the accounts?
Average Clustering Coefficient	How strong is the clique formation in the graph?
Highest Centrality Account	Which account is particularly central to the communication?

The *density* of a graph can be interpreted as the degree of interconnection of the nodes. In the case of a communication graph, a structural change in density indicates a change in the communication behavior of the corresponding e-mail accounts. In the context of a police investigation, it is assumed that such a change is suspicious and e-mails in the corresponding period should be examined more closely.

The *average clustering coefficient* of a graph describes how strongly nodes tend to form cliques. A clique is a set of nodes, all of which are interconnected. Studies such as [15] show that individuals tend to communicate particularly intensely within their close network in response to unusual externalities. It is assumed that changes in clique formation behavior may reveal such externalities to investigators.

Centrality metrics can be used to evaluate the importance of a node within a graph. The *Highest Centrality Account* is the node, which takes the maximum centrality. In a communication graph, the centrality of a node can provide information about how important a certain actor is within the graph. For example, one possible interpretation of an important actor may be that it controls or significantly influences the flow of information in a group. Accordingly, a change of the node with the highest centrality can be an indicator for structural changes within a dynamic communication graph. For example, in the context of police investigations of criminal networks, such a change may indicate a shift in the hierarchy of the network. Since there are different ways to evaluate the centrality of nodes, various metrics can be tested using this approach. Exemplary metrics are *degree centrality* and *betweenness centrality*.

At this stage, even a simple visualizing of the time series can already identify time points and periods of particular interest, for example times of particularly low or intensive communication, or periodicity in the communication. However, we propose an approach to automatically detect structural changes in the following step.

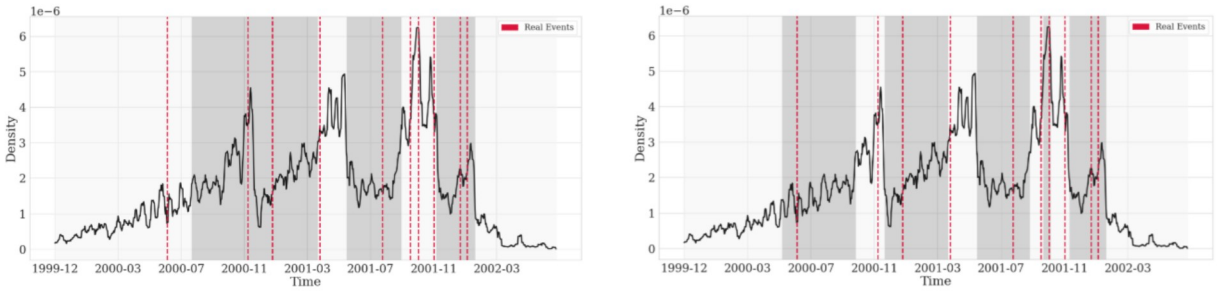
(4) Identification of anomalies

For the automated identification of structural changes in the time-series of a graph metric we propose change-point-detection (CPD). CPD algorithms divide a time series in sub-segments that are as homogeneous as possible. The boundaries between these segments are called change points. After determining these change points, investigators can examine the e-mail communication around the corresponding time in order to detect interesting events.

To design a CPD, we follow the approach of [21]. According to this approach, a CPD consists of a search strategy, a cost function and, if known, the number of change points to be determined. If the number of change points is known, this approach uses the algorithm of [22] based on dynamic programming. As a search strategy when the number of change points is unknown, [21] proposes the Pruned Exact Linear Time (PELT) algorithm of [23]. In this case, a linear penalty function $p_{linear}(k) = \beta|k|$ is used to limit the number of change points k. The penalty coefficient β is calculated depending on the cost function. The cost function is part of a CPD algorithm and chosen depending on the change to be detected. Typical cost functions are change-in-mean or change-in-variance (and others). We suggest change-in-mean for our application, so we use the cost function in Equation 1.

$$c_\mu(y_{a..b}) = \sum_{t=a+1}^{b} \|y_t - \bar{y}_{a..b}\|_2^2 \tag{1}$$

Hereby, $y_{a..b}$ are sub-segments of variable length of the time-series under investigation $y = \{y_t\}_{t=1}^{T}$.



(a) Unknown no. of change points

(b) Known no. of change points

Fig. 3: Change points of the density of the dynamic graph of Enron e-mail communication. The identified change points are located at the transition of the colors.

4 Case Study

We test our method on the famous Enron e-mail data set [4]. As part of the data cleaning process, e-mails sent to multiple recipients are transformed into multiple e-mails, each with a single recipient. After cleaning, the data set consists of 1,115,923 e-mails during the period from December 1999 to July 2002. A period of seven days per snapshot is chosen to model the dynamic graph. On the dynamic graph model, the previously mentioned graph metrics are calculated for each snapshot. Then, the resulting time series are segmented by the described approach to CPD using the `ruptures` library in Python 3 of [21]. The segment boundaries are interpreted as change points. The automatically determined change points are compared with real events from the environment of the Enron

Table 3: Real events from the environment of the Enron Group.

Date (dd.mm.yyyy)	Event
01.07.2000	Cooperation with Blockbuster
01.12.2000	Announcement of Jeffrey Skilling taking over as CEO
17.01.2001	Start of the electricity crisis in California
17.04.2001	*Asshole* call
14.08.2001	Skilling resigns / Kenneth Lay takes over as CEO
16.10.2001	Publication of high losses in the last quarter
22.-31.10.2001	First investigations against Enron become public
20.-26.11.2001	Drastic drop in Enron share price
09.01.2002	Announcement of investigations into accounting fraud
23.01.2002	Kenneth Lay resigns as CEO

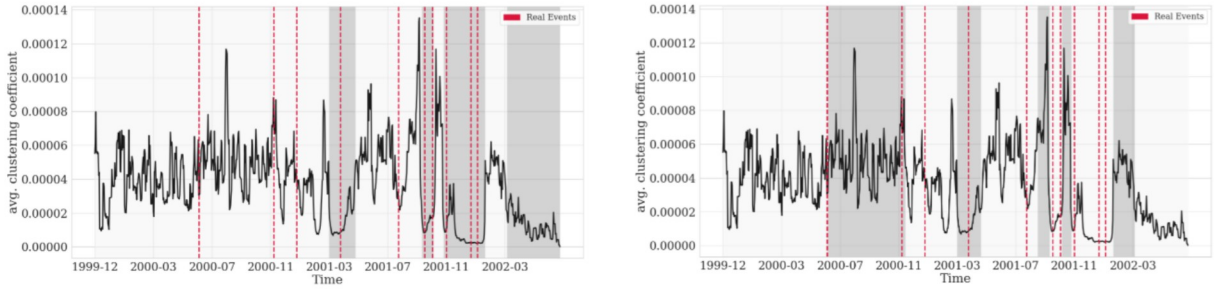
Group in the period under investigation showed in Table 3. These events are taken from chronicles of the Enron scandal from an article in a German business magazine [24] and a website of the UMKC School of Law [25].

In the following we discuss the analysis of the *density* and *average clustering coefficient* for the ENRON dataset. We use the PELT algorithm and chose a penalty coefficient $\beta = max(y_t)^2$, where y_t describes the time series of the density. In an intuitive sense, this penalty coefficient means that an additional change point will penalize the value of the cost function as much as the highest value of the time series.

In our figures, which we introduce in the following, we use two different grey-scale colors to mark the identified segments. The identified change points are consequently located at the transition of the colors. The real events cited in the table above are market by dotted vertical lines.

Density

Figure 3a shows the automatically identified change points on the time series of the density by the PELT algorithm. Six change points are identified by the PELT algorithm. Since the algorithm divides the time series into sub-segments that are as homogeneous as possible, the change points do not necessarily correspond to points in time that are particularly conspicuous visually. The plot shows that one real event in April 2001 corresponds exactly to one change point. Three other change points are close to other real events that are in the period of increased density at the end of 2001. The remaining real events cannot be predicted by this approach. Due to the unknown number of change points, obviously not each change point can be assigned to each real event.



(a) Unknown no. of change points

(b) Known no. of change points

Fig. 4: Change points of the average clustering coefficient of the dynamic graph of Enron e-mail communication. The identified change points are located at the transition of the colors.

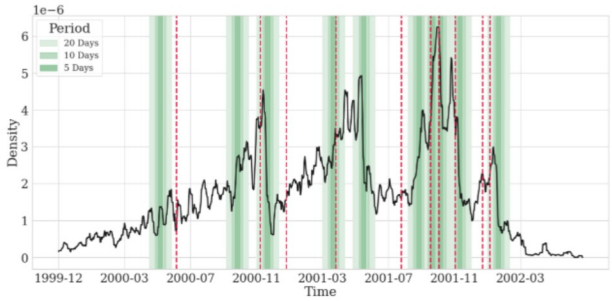


Fig. 5: Curve of the density of the dynamic graph of the Enron data set with real events and periods of investigation of 5, 10 and 20 days before and after ten change points

We also fix the number of change points and use [22]. It shall be tested how close the proposed method does predict real events. Therefore, according to the number of real events, the number of change points searched is set to ten[1]. The change points determined in this way are visualized in Figure 3b. As with the testing of the first approach, it is observed from Figure 3b that some change points are close to real events. In addition to the change points identified by the PELT algorithm, change points in late 2000 and late 2001 can be identified that are close to real events. Some other real events can again not be predicted (however, these events may not even be contained in the communication structure).

Average Clustering Coefficient

The analysis of the time series of the average clustering coefficient yields a similar result. By the PELT algorithm, illustrated in Figure 4a, six change points are identified. Three of them are very close to real events at the end of 2001. By specifying a searched number of change points, shown in Figure 4b, a total of five real events can be identified with sufficient precision. Also in this case, two further events in July 2001 and January 2002 are located near change points.

Recommendation for investigators

When applying the method presented here as part of an investigation, it is recommended that e-mails from periods before and after each change point are examined in detail, as these periods contain possible interesting content for investigations. This is why an exact determination of the real event is not mandatory. However, when narrowing down the time period for a detailed investigation of e-mails around a change point, it is necessary to balance between the least possible effort for the investigator and the thoroughness of the investigation.

Figure 5 visualizes the curve of the *density* of the dynamic graph. As in the previous plots, the real events are marked by dotted lines. In addition, in this representation, the ten determined change points are color-coded by their periods of investigation of lengths 5, 10 and 20 days. Some change points are close enough to each other that their observation periods may even overlap. Especially in the period from August to December 2001, some time periods already overlap for a threshold of 10 days before and after each change point, while this can be observed for 20 days in several cases. Because of these overlaps, there is no need to examine e-mail communications of, say, 20 days

[1] To identify ten change points, the time series must be divided into eleven segments as the end of the last segment is not considered a change point

before and after each change point. In this example, if we compare the proportion of days in the total time horizon and the share of e-mails that must be examined by the corresponding observation periods, the proportions are as shown in Table 4. This shows that in the example, 30 % of the real events are already covered from an examination

Table 4: Proportion of time periods and e-mails to be investigated in the total observation horizon as a percentage of the observation period per change point

Period under investigation	share time period	share e-mails	Precision
+/- 5 Days	12 %	20 %	30 %
+/- 10 Days	21 %	37 %	40 %
+/- 20 Days	38 %	59 %	60 %

of only 12 % of the entire time horizon. These may be periods of particularly intensive communication, during which an above-average number of e-mails must be examined. Still, however, for 30 % of real events only 20 % of e-mails must be examined. This speeds up the investigation and can lead to more evidence being seized more quickly. These can speed up the entire investigation process.

The decisive added value of this approach for investigators thus lies in the identification of change points that mark important structural shifts in a communication network. By prioritizing the investigation of e-mails around the corresponding points in time, scarce capacities can thus be used efficiently.

5 Conclusion

We propose an algorithm for automatic detection of important events in criminal cases using only e-mail metadata. The method is based on a graph representation of the data and an automated detection of change points on time-series of graph metrics. Using the Enron data set, the case study shows that structural anomalies in a dynamic graph are indicative of real events. If an investigator is provided with the identified change points during the exploration of an e-mail data set, she can examine the e-mail communication around the identified time periods in detail. This decisively narrows down the volume of e-mails to be examined and capacities are freed up. Since the method developed here is only tested on a single data set as part of the case study, no general statement can be made about the potential of this method in police investigations. This requires testing by police investigators and testing on additional e-mail data sets. It should also be examined how the method behaves on larger or smaller data sets. In addition, the comparison of subjectively selected, real events and automatically determined change points does not represent a reliable form of evaluation that allows a statement about the applicability of this method for the unsupervised exploration of e-mail data.

Regardless of the automatically determined change points, insights such as time periods of special interest for further investigations can also be drawn from the visualization of the time-series of graph metrics. For this purpose, the investigation of further graph metrics can be part of future research, in particular graph metrics on node level. Also different cost function in the CPD can be considered. Due to the unsupervised nature of the problem it may be useful to implement an ensemble of parameter variations in practical application.

References

1. Bavarian State Ministry of the Interior, f.S., Integration: Betrug im Gesundheitswesen (2018)
2. SueddeutscheZeitung: Millionenbetrug bei pflegediensten in münchen und augsburg. (2021)
3. Federal Ministry of Education and Research: Bekämpfung von Abrechnungsbetrug und Korruption im Gesundheitswesen (Kriminelle Netzwerke) (2021)
4. Cohen, W.: Enron email dataset (2015)
5. Stolfo, S.J., Hershkop, S., Wang, K., Nimeskern, O., Hu, C.W.: Behavior profiling of email. In Chen, H., ed.: Intelligence and security informatics. Volume 2665 of Lecture Notes in Computer Science. Springer, Berlin and Heidelberg (2003) 74–90
6. Stolfo, S.J., Li, W.J., Hershkop, S., Wang, K., Hu, C.W., Nimeskern, O.: Detecting viral propagations using email behavior profiles. ACM transactions on internet technology (TOIT) (2004) 128–132
7. Stolfo, S.J., Hershkop, S.: Email mining toolkit supporting law enforcement forensic analyses. In: Proceedings of the 2005 national conference on Digital government research. (2005) 221–222
8. Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., Benredjem, D.: Towards an integrated e-mail forensic analysis framework. Digital Investigation **5**(3-4) (2009) 124–137
9. Meng, F., Wu, S., Yang, J., Yu, G.: Research of an e-mail forensic and analysis system based on visualization. In: 2009 Asia-Pacific Conference on Computational Intelligence and Industrial Applications, Piscataway, NJ, IEEE (2009) 281–284

10. Sobiya R. Khan, Smita M. Nirkhi, R. V. Dharaskar: E-mail data analysis for application to cyber forensic investigation using data mining. IJAIS Proceedings on 2nd National Conference on Innovative Paradigms in Engineering and Technology (NCIPET 2013) **NCIPET**(3) (2013) 1–4
11. Haggerty, J., Haggerty, S., Taylor, M.: Forensic triage of email network narratives through visualisation. Information Management & Computer Security **22**(4) (2014) 358–370
12. Stadlinger, J., Dewald, A.: A forensic email analysis tool using dynamic visualization. Journal of Digital Forensics, Security and Law (2017)
13. Kossinets, G., Watts, D.J.: Empirical analysis of an evolving social network. Science (New York, N.Y.) **311**(5757) (2006) 88–90
14. Kondor, D., Csabai, I., Szüle, J., Pósfai, M., Vattay, G.: Inferring the interplay between network structure and market effects in bitcoin. New Journal of Physics **16**(12) (2014) 125003
15. Romero, D.M., Uzzi, B., Kleinberg, J.: Social networks under stress. (2016)
16. Miller, H., Mokryn, O.: Size agnostic change point detection framework for evolving networks
17. Peel, L., Clauset, A.: Detecting change points in the large-scale structure of evolving networks. Proc. of the 29th International Conference on Artificial Intelligence (AAAI) (2015)
18. Huang, S., Hitti, Y., Rabusseau, G., Rabbany, R.: Laplacian change point detection for dynamic graphs. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, New York, NY, USA, ACM (2020) 349–358
19. Zhu, T., Li, P., Yu, L., Chen, K., Chen, Y.: Change point detection in dynamic networks based on community identification. IEEE Transactions on Network Science and Engineering **7**(3) (2020) 2067–2077
20. Kendrick, L., Musial, K., Gabrys, B.: Change point detection in social networks—critical review with experiments. Computer Science Review **29** (2018) 1–13
21. Truong, C., Oudre, L., Vayatis, N.: Selective review of offline change point detection methods. Signal Processing **167**(4) (2020) 107299
22. Bellman, R.: On a routing problem. Quarterly of Applied Mathematics **16**(1) (1958) 87–90
23. Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. Journal of the American Statistical Association **107**(500) (2012) 1590–1598
24. Frentz, C.: Enron: Chronik einer Rekord-Pleite. Manager Magazin **2003** (2003)
25. Linder, D.: The enron trial: A chronology (2022)