Saarland University
Faculty for Mathematics and Computer Science
Department of Computer Science

Novel Approaches to Anonymity and Privacy in Decentralized, Open Settings

Dissertation
zur Erlangung des Grades
des Doktors der Ingenieurwissenschaften
der Fakultät für Mathematik und Informatik
der Universität des Saarlandes

von
Praveen Manoharan

Saarbrücken,
April 2019

Tag des Kolloquiums: 29.03.2019

Dekan: Prof. Dr. Sebastian Hack

Prüfungsausschuss:

Vorsitzender:	Prof. Bernd Finkbeiner, Ph.D.
Berichterstattende:	Prof. Dr. Michael Backes
	Prof. Dr. Krishna Gummadi
Akademischer Mitarbeiter:	Dr. Yang Zhang

Zusammenfassung

Das Internet hat in den letzten zwei Jahrzehnten eine drastische Transformation erlebt und entwickelte sich dabei von einem einfachen Kommunikationsnetzwerk zu einer globalen Multimedia Plattform auf der Milliarden von Nutzern aktiv Informationen austauschen. Diese Transformation hat zwar einen gewaltigen Nutzen und vielfältige Vorteile für die Gesellschaft mit sich gebracht, hat aber gleichzeitig auch neue Herausforderungen und Gefahren für online Privacy mit sich gebracht mit der die aktuelle Technologie nicht mithalten kann.

In dieser Dissertation präsentieren wir zwei neue Ansätze für Anonymität und Privacy in dezentralisierten und offenen Systemen. Mit unserem ersten Ansatz untersuchen wir das Problem der Attribut- und Identitätspreisgabe in offenen Netzwerken und entwickeln hierzu den Begriff der (k, d)-Anonymität für offene Systeme welchen wir extensiv analysieren und anschließend experimentell validieren. Zusätzlich untersuchen wir die Beziehung zwischen Anonymität und Unlinkability in offenen Systemen mithilfe des Begriff der (k, d)-Anonymität und zeigen, dass, im Gegensatz zu traditionell betrachteten, abgeschlossenen Systeme, Anonymität innerhalb einer Online Community nicht zwingend die Unlinkability zwischen verschiedenen Online Communitys impliziert.

Mit unserem zweiten Ansatz untersuchen wir die transitive Diffusion von Information die in Sozialen Netzwerken geteilt wird und sich dann durch die paarweisen Interaktionen von Nutzern durch eben dieses Netzwerk ausbreitet. Wir entwickeln eine neue Methode zur Kontrolle der Ausbreitung dieser Information durch die Minimierung ihrer Exposure, was dem Besitzer dieser Information erlaubt zu kontrollieren wie weit sich deren Information ausbreitet indem diese initial mit einer sorgfältig gewählten Menge von Nutzern geteilt wird. Wir implementieren die hierzu entwickelten Algorithmen und untersuchen die praktischen Grenzen der Exposure Minimierung, wenn sie von Nutzerseite für große Netzwerke ausgeführt werden soll.

Beide hier vorgestellten Ansätze verbindet eine Neuausrichtung der Aussagen die diese bezüglich Privacy treffen: wir bewegen uns weg von beweisbaren Privacy Garantien für abgeschlossene Systeme, und machen einen Schritt zu robusten Privacy Risikoeinschätzungen für dezentralisierte, offene Systeme in denen solche beweisbaren Garantien nicht möglich sind.

Abstract

The Internet has undergone dramatic changes in the last two decades, evolving from a mere communication network to a global multimedia platform in which billions of users actively exchange information. While this transformation has brought tremendous benefits to society, it has also created new threats to online privacy that existing technology is failing to keep pace with.

In this dissertation, we present the results of two lines of research that developed two novel approaches to anonymity and privacy in decentralized, open settings. First, we examine the issue of attribute and identity disclosure in open settings and develop the novel notion of (k, d)-anonymity for open settings that we extensively study and validate experimentally. Furthermore, we investigate the relationship between anonymity and linkability using the notion of (k, d)-anonymity and show that, in contrast to the traditional closed setting, anonymity within one online community does necessarily imply unlinkability across different online communities in the decentralized, open setting.

Secondly, we consider the transitive diffusion of information that is shared in social networks and spread through pairwise interactions of user connected in this social network. We develop the novel approach of exposure minimization to control the diffusion of information within an open network, allowing the owner to minimize its exposure by suitably choosing who they share their information with. We implement our algorithms and investigate the practical limitations of user side exposure minimization in large social networks.

At their core, both of these approaches present a departure from the provable privacy guarantees that we can achieve in closed settings and a step towards sound assessments of privacy risks in decentralized, open settings.

Background of this Dissertation

This dissertation is based on the papers summarized in the following. The author of this dissertation (henceforth simply noted as *the author*) contributed to all of these papers as the main author.

The author had the idea for and is primarily responsible for the development of the d-convergence anonymity framework and its experimental validation presented in the first part of [P2]. Pascal Berrang assisted with the execution of the evaluations and is primarily responsible for the development of its second part on authorship attribution (that is not presented in this dissertation). All authors performed reviews of the paper.

Motivated by the experimental results in [P2], the author lead further efforts for extensive experimental evaluations of the (k, d)-anonymity notion developed in [P2], and its relation to linkability in [P1]. Pascal Berrang again assisted with the execution of the evaluations. He, together with Oana Goga and Krishna Gummadi, assisted in the interpretation of the experimental results. The interpretation that was finally presented in the publication was then finalized by the author of this dissertation. All authors also performed reviews of the paper.

For [P3], the author developed the idea of leveraging influence minimization in diffusion networks to allow users to control the spread of their shared information in decentralized social networks. Manuel Gomez-Rodriguez assisted with his expert knowledge on continuous-time diffusion networks and influence, while the author developed the notions of privacy policies for information diffusion and the corresponding exposure minimization optimization problem. With Manuel's background knowledge on the related influence maximization problems and submodular set functions, the author also primarily developed the algorithmic approximations to the exposure minimization problem. Again, all authors performed reviews of the paper.

As follow up to the above work, the author presents an implementation of the approximation algorithm developed in [P3] and discusses the various challenges of exposure minimization in practice in [P4]. The authors solved one of main issues that is the scalable estimation of exposure by adopting a previously known algorithm for scalable influence estimation, and performed extensive evaluations to validate the performance of the presented algorithm. All authors performed reviews of the paper.

[P1] Backes, M., Berrang, P., Goga, O., Gummadi, K. P., and Manoharan, P. On Profile Linkability Despite Anonymity in Social Media Systems. In: *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. ACM. 2016, 25–35.

[P2] Backes, M., Berrang, P., and Manoharan, P. From Zoos to Safaris—From Closed-World Enforcement to Open-World Assessment of Privacy. In: *Foundations of Security Analysis and Design VIII*. Springer, 2015, 87–138.

[P3] Backes, M., Gomez-Rodriguez, M., Manoharan, P., and Surma, B. Reconciling Privacy and Utility in Continuous-time Diffusion Networks. In: *Proceedings of the 30th Computer Security Foundations Symposium (CSF)*. IEEE. 2017, 292–304.

[P4] Manoharan, P. and Backes, M. XpoMin: Towards Practical Exposure Minimization in Continuous Time Diffusion Networks. In: *under submission*. 2018.

Further Contributions of the Author

[S1] Backes, M., Berrang, P., Humbert, M., and Manoharan, P. Membership privacy in MicroRNA-based studies. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, 319–330.

[S2] Backes, M., Kate, A., Manoharan, P., Meiser, S., and Mohammadi, E. AnoA: A framework for analyzing anonymous communication protocols. In: *Proceedings of the 26th Computer Security Foundations Symposium (CSF)*. IEEE. 2013, 163–178.

[S3] Backes, M., Manoharan, P., and Mohammadi, E. TUC: Time-sensitive and Modular Analysis of Anonymous Communication. In: *Proceedings of the 27th Computer Security Foundations Symposium (CSF)*. IEEE. 2014, 383–397.

[S4] Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* (2017).

[S5] Grosse, K., Papernot, N., Manoharan, P., Backes, M., and McDaniel, P. Adversarial Examples for Malware Detection. In: *Proceedings of the 22nd European Symposium on Research in Computer Security (ESORICS)*. Springer. 2017, 62–79.

Contents

List of Figures

List of Tables

List of Algorithms

1 Introduction

The Internet has undergone dramatic changes in the last two decades, evolving from a mere communication network to a global multimedia platform in which billions of users not only actively exchange information, but increasingly conduct sizable parts of their daily lives. While this transformation has brought tremendous benefits to society, it has also created new threats to online privacy that existing technology is failing to keep pace with. Users tend to reveal personal information without considering the widespread, easy accessibility, potential linkage and permanent nature of online data. Many cases reported in the press show the resulting risks, which range from public embarrassment and loss of prospective opportunities (e.g., when applying for jobs or insurance), to personal safety and property risks (e.g., when sexual offenders or burglars learn users' whereabouts online). The resulting privacy awareness and privacy concerns of Internet users have been further amplified by the advent of the Big-Data paradigm and the aligned business models of personalized tracking and monetizing personal information in an unprecedented manner.

Developing a suitable methodology to reason about the privacy of users in such a large-scale, open web setting, as well as corresponding tool support in the next step, requires at its core formal privacy notions that live up to the now increasingly dynamic dissemination of unstructured, heterogeneous user content on the Internet: While users traditionally shared information mostly using public profiles with static information about themselves, nowadays they disseminate personal information in an unstructured, highly dynamic manner, through content they create and share (such as blog entries, user comments, a "Like" on Facebook), or through the people they befriend or follow.

Furthermore, ubiquitously available background knowledge about a dedicated user needs to be appropriately reflected within the model and its reasoning tasks, as it can decrease a user's privacy by, for instance, allowing the adversary to to infer further sensitive information. As an example, Machine Learning and other Information Retrieval techniques provide comprehensive approaches for profiling a user's actions across multiple Online Social Networks, up to a unique identification of a given user's profiles for each such network.

In this dissertation, we present the results of two lines of research that developed two novel approaches to anonymity and privacy in decentralized, open settings. First, we closely examined the issue of attribute and identity disclosure in open settings. This issue most closely resembles the database privacy settings traditionally considered in the literature. We will show that the traditional privacy notions cannot be suitably applied to open settings, and develop a novel notion of anonymity for open settings that we extensively study.

Secondly, we consider the transitive diffusion of information that is shared in social networks and spread through pairwise interactions of user connected in this social network. In such a setting, once information is shared, the owner of this information quickly loses control over who sees the information. This sets a striking contrast to access control solutions in closed or centralized settings where access to information is centrally regulated. We develop a novel approach to controlling the diffusion of information within a network, allowing the owner to minimize its exposure by suitably choosing who they share their information with.

At their core, both of the presented approaches are driven by the central idea that

in decentralized, open settings, provable guarantees of privacy seem to be impossible. Instead, we follow the idea of assessing the risk of privacy violations, and thus assisting users to perform informed decisions with respect to their information sharing behavior.

Our work on anonymity and privacy in decentralized, open settings stretches across the following publications which each contributed to the development of the two novel approaches presented in this dissertation:

d-convergence In this work [P2], we investigated the central challenges of attribute and identity disclosure in open settings and compare it to the privacy settings traditionally considered in the literature. From our observations, we then constructed a novel privacy framework in which we formalize the unstructured and heterogeneous dissemination of information across various platforms and quantify the risk of information disclosure against strong adversaries. We showed that, in this framework, traditional privacy notions do not provide meaningful privacy guarantees, and therefore highlight the need for new anonymity and privacy notions that tackle the challenges of open settings.

As a solution, we then developed the notion of d-convergence and (k, d)-anonymity, which is a generalization of the traditional k-anonymity notion to open settings. In this generalization we relaxed the requirement for identity found in the traditional k-anonymity notion to requiring a certain degree of similarity between profiles within an anonymity set/equivalence class. We finally validated that (k, d)-anonymity indeed quantifies the anonymity of users within online social networks and thus presented a first approach to anonymity in decentralized, open settings.

Anonymity and Linkability in Open Settings As follow up work to the above publication, we investigated the relation between anonymity and linkability in open settings in [P1]. With extensive evaluations on the Reddit social media platform, we showed that, in contrast to traditional privacy settings, anonymity alone does not automatically imply unlinkability in decentralized, open settings. For instance, in a quantitative evaluation of linking profiles across communities in Reddit, just using the size of the anonymous subsets of an identity underestimates its linkability risks in more than 40% of the cases.

To produce a better linkability estimate, we then proposed the notion of local matching sets that combines the anonymity set of a target identity with further information about the source identity that the adversary tries to link to the target identity. Our evaluations show that the size of the local matching set much more accurately estimates linkability risks than just considering anonymity alone.

These evaluations show, as also modeled by the privacy framework developed in [P2], that the unstructured dissemination of information can allow some identifying attributes to remain even if a user achieves a certain degree of anonymity within a community. The identifying attributes then consequently allow for the linking of this user's profiles across online platforms. To obtain accurate linkability risk estimates it is therefore important to also take into account which identities the adversary is trying to match.

Reconciling Utility and Privacy in Diffusion Networks Orthogonal to the issue of attribute and identity disclosure due to shared information is the issue of controlling who actually sees information shared by a user. While in traditional, closed settings there is usually centralized mechanisms that allows for the enforcement of various access control mechanisms, in decentralized, open settings the control over information is quickly lost once shared. In [P3], we investigate the issue of information diffusion in social networks, and propose the exposure minimization approach to controlling the spread of information: by suitably choosing the nodes with which we share a piece of information initially, we can minimize the expected number of (malicious) nodes in the network that will learn the shared information within a certain time frame despite potential transitive diffusion of information. We showed that the corresponding optimization problems are NP-hard, but can be, by leveraging the submodularity of the objective function, efficiently approximated to a problem instance specific constant.

XpoMin As follow up work to the above publication, we the investigated the issue of exposure minimization in practice and implemented the necessary exposure estimation and minimization algorithms. The resulting architecture allows for a separation of the required computations in various computation phases, with ultimately a near constant computation time for the actual minimization during run-time. While the necessary pre-computations still remain high, and therefore nonviable to be computed on user side, we still presented a scalable approach to exposure minimization in social networks, that can a) either be offered as a service by social network providers, or b) used as a baseline for future improvements.

Outline

We begin by discussing the challenges of anonymity and privacy in decentralized, open settings in Chapter 2. We then develop the d-convergence privacy framework in Chapter 4, and evaluate the relation between anonymity and linkability in open settings in Chapter 5. In Chapter 6, we introduce the problem of privacy in diffusion networks, and develop our exposure minimization approach to this issue. The corresponding implementation of the exposure estimation and minimization algorithms, and their evaluation, is presented in Chapter 7. Finally, we conclude this dissertation in Chapter 8.

2

Challenges of Anonymity and Privacy in Decentralized, Open Settings

Before we delve into the technical parts of this dissertation, we first discuss what actually constitutes a decentralized, open settings in contrast to the closed settings traditionally considered in the literature. Then, we investigate two specific use cases of privacy in decentralized, open settings: first, we take a look at the issue of identity disclosure, i.e. the linking of a persons real identity to information disseminated through pseudonyms in social networks. Second, we consider the issue of controlling the spread of shared information in connected systems. Using examples, we first illustrate the challenges corresponding to these use cases, and then argue why traditional solutions do not solve these challenges.

2.1 The Decentralized and Open Setting

In the last two decades, the Internet has fundamentally changed how many user consume media and information. Instead of solely being recipients of services provided by third party providers, users have transformed two simultaneously function as producers in addition to consumers: social media platforms such as Facebook, Twitter and Reddit allow users to share media and information with other users, and also to further transitively propagate the learned information to other users. Most markedly, this information sharing additionally happens without regulation from any (trusted) third party.

In this work, we summarize these circumstances under the notion of *decentralized, open settings*. Under this notion, we capture the following properties of online information sharing and social media platforms:

1. The communication in the network is *decentralized*, i.e. communication happens directly between two or more users of the communication platform and is not regulated by any (trusted) third party with elevated privileges. Furthermore, all users in the network can potentially function as sources and recipients of information simultaneously.

2. The communication network itself is *open*, i.e. there is pre-defined set of users that participate in the communication and no specific structure to which shared information has to adhere to.

These properties present a significant contrast to the closed settings that are traditionally considered in the privacy literature (cf. Chapter 3 for a discussion of related work). In these closed settings, there typically is a central information source (e.g. a database and the database curator that allows access to this database) in which information from a pre-defined set of users is available in a very specific data format. The access to this information is then regulated and curated in such a way such that traditional privacy mechanisms such as k-anonymity and Differential Privacy can achieve provable privacy guarantees.

In the following, we take a closer look at the two issues of information disclosure and controlling access to information in decentralized open settings, identify the corresponding challenges and discuss why traditional solutions do not apply in the decentralized, open setting.

2.2 Information Disclosure in Decentralized, Open Settings

We first investigate the use case of information disclosure in decentralized, open settings. Here, we are interested in the likelihood that an adversary can, given a collection of information, infer the identity of the user to whom this information belongs to, or, alternatively, infer additional (sensitive) information about this users using, for instance, additional background knowledge. This would naturally violate the user's privacy since he only intended to release the original collection of information without either revealing their information or any other sensitive information about themselves.

2.2.1 Example: Identity Disclosure

First, consider the following example: Employer Alice receives an application by potential employee Bob which contains personal information about Bob. Before she makes the decision on the employment of Bob, however, she searches the <internet and tries to learn even more about her potential employee. A prime source of information are, for example, Online Social Networks (OSNs) which Alice can browse through. If she manages to identify Bob's profile in such an OSN she can then learn more about Bob by examining the publicly available information of this profile.

In order to correctly identify Bob's profile in an OSN, Alice takes the following approach: based on the information found in Bob's application, she constructs a model θ_B that contains all attributes, such as name, education or job history, extracted from Bob's application. She then compares this model θ_B to the profiles $P_1, \ldots, P_n$ found in the OSNs and ranks them by similarity to the model θ_B. Profiles that show sufficient similarity to the model θ_B are then chosen by Alice as belonging to Bob. After identifying the (for Alice) correctly matching profiles $P_1^*, \ldots, P_i^*$ of Bob, Alice can finally merge their models $\theta_1^*, \ldots, \theta_i^*$ with θ_B to increase her knowledge about Bob.

Bob now faces the problem that Alice could learn information about him that he does not want her to learn. He basically has two options: he either does not share this critical information at all, or makes sure that his profile is not identifiable as his. In OSNs such as Facebook, where users are required to identify themselves, Bob can only use the first option. In anonymous or pseudonymous OSNs such as Reddit or Twitter, however, he can make use of the second option. He then has to make sure that he does not share enough information on his pseudonymous profiles that would allow Alice to link his pseudonymous profile to him personally.

In this work, we are mostly concerned with the second option: we cannot protect an entity ϵ against sharing personal information through a profile which is already uniquely identified with the entity ϵ. We can, however, estimate how well an pseudonymous account of ϵ can be linked to ϵ, and through this link, learn personal information about ϵ. As the example above shows, we can essentially measure privacy in terms of similarity of an entity ϵ in a collection of entities $\mathcal{E}$.

The identifiability of ϵ then substantially depends on the attributes ϵ exhibits in the context of $\mathcal{E}$ and does not necessarily follow the concept of personally identifiable information (PII) as known in the more common understanding of privacy and in privacy and data-protection legislation [26]: here, privacy protection only goes as far as protecting this so-called personally identifiable information, which often is either

not exactly defined, or restricted to an a-priori-defined set of attributes such as name, Social Security number, etc. We, along with other authors in the literature [81, 80], find however that the set of critical attributes that need to be protected differ from entity to entity, and from community to community. For example, in a community in which all entities have the name "Bob", exposing your name does not expose any information about yourself. In a different community, however, where everyone has a different name, exposing your name exposes a lot of information about yourself.

In terms of the privacy taxonomy formulated by Zheleva and Getoor [118], the problem we face corresponds to the identity disclosure problem, where one tries to identify whether and how an identity is represented in an OSN. We think that this is one of the main concerns of users of frequently used OSNs, in particular those that allow for pseudonymous interactions: users are able to freely express their opinions in these environments, assuming that their opinions cannot be connected to their real identity. However, any piece of information they share in their interactions can leak personal information that can lead to identity disclosure, defeating the purpose of such pseudonymous services.

To successfully reason about the potential disclosure of sensitive information in such open settings, we first have to consider various challenges that have not been considered in traditional privacy research. After presenting these challenges, we discuss the implications of these challenges on some of the existing privacy notions, before we consider other relevant related work in the field.

2.2.2 Challenges of Privacy in Open Settings

In this subsection, we introduce the challenges induced by talking about privacy in open settings:

C1) Modeling heterogeneous information. We require an information model that allows for modeling various types of information and that reflects the heterogeneous information shared throughout the Internet. This models needs to adequately represent personal information that can be inferred from various sources, such as static profile information or from user-generated content, and should allow statistical assessments about the user, as is usually provided by knowledge inference engines. We propose a solution to this challenge in Section 4.4.1.

C2) User-specified privacy requirements. We have to be able to formalize user-specified privacy requirements. This formalization should use the previously mentioned information model to be able to cope with heterogeneous information, and specify which information should be protected from being publicly disseminated. We present a formalization of user privacy requirements in Section 4.4.4.

C3) Information sensitivity. In open settings, information sensitivity is is a function of user expectations and context: we therefore need to provide new definitions for sensitive information that takes user privacy requirements into account. We present context- and user-specific definitions of information sensitivity in Section 4.4.5.

C4) Adversarial knowledge estimation. To adequately reason about disclosure risks in open settings we also require a parameterized adversary model that we can instantiate

with various assumptions on the adversary's knowledge: this knowledge should include the information disseminated by the user, as well as background knowledge to infer additional information about the user. In Section 4.4, we define our adversary model based on statistical inference.

In Chapter 4, we provide a rigorous formalization of these requirements, leading to a formal framework for information disclosure in open settings, and will furhter instantiate this framework to reason about the identity disclosure in particular. In Chapter 5 we then investigate the relationship between the notions of anonymity and linkability in decentralized, open settings.

2.2.3 Inadequacy of Existing Models

Common existing privacy notions such as k-anonymity [103], l-diversity [74], t-closeness [70] and the currently most popular notion of Differential Privacy [30] provide the technical means for privacy-friendly data-publishing in a closed-world setting: They target scenarios in which all data is available from the beginning, from a single data source, remains static and is globally sanitized in order to provide rigorous privacy guarantees. In what follows, we describe how these notions fail to adequately address the challenges of privacy in open settings discussed above.

a) Absence of structure and classification of data. All the aforementioned privacy models require an a-priori structure and classification of the data under consideration. Any information gathered about an individual thus has to be embedded in this structure, or it cannot be seamlessly integrated in these models.

b) No differentiation of attributes. All of these models except for Differential Privacy require an additional differentiation between key attributes that identify an individual record, and sensitive attributes that a users seeks to protect. This again contradicts the absence of an a-priori, static structure in our setting. Moreover, as pointed out above and in the literature [81], such a differentiation cannot be made a-priori in general, and it would be highly context-sensitive in the open web setting.

c) Ubiquitously available background knowledge. All of these models, except for Differential Privacy, do not take into account adversaries that utilize ubiquitously available background knowledge about a target user to infer additional sensitive information. A common example of background knowledge is openly available statistical information that allows the adversary to infer additional information about an identity.

d) Privacy for individual users. All these models provide privacy for the whole dataset, which clearly implies privacy of every single user. One of the major challenges in open settings such as the Internet, however, is that accessing and sanitizing all available data is impossible. This leads to the requirement to design a local privacy notion that provides a lower privacy bound for every individual user, even if we only have partial access to the available data.

The notion of Differential Privacy only fails to address some of the aforementioned requirements (parts *a* and *d*), but it comes with the additional assumption that the adversary knows almost everything about the data set in question (everything except for the information in one database entry). This assumption enables Differential Privacy

to avoid differentiation between key attributes and sensitive attributes. This strong adversarial model, however, implies that privacy guarantees are only achievable if the considered data is globally perturbed [25, 31, 32], which is not possible in open web settings.

The conceptual reason for the inadequacy of existing models for reasoning about privacy in open web settings is mostly their design goal: Privacy models have thus far mainly been concerned with the problem of attribute disclosure within a single data source: protection against identity disclosure was then attempted by preventing the disclosure of any (sensitive) attributes of a user to the public. In contrast to static settings such as private data publishing, where we can decide which information will be disclosed to the adversary, protection against any attribute disclosure in open settings creates a very different set of challenges (as discussed above).

2.3 Information Exposure in Decentralized, Open Settings

We next investigate the use case of controlling the exposure of information, i.e. controlling the expected number of (malicious) users in a network that receive a piece of information that was originally shared with a small set of users. Due to the transitive diffusion of this information by, for instance, user re-sharing information they have learned to other users in the network, the shared information can quickly escape the original user's control.

2.3.1 Example: Exposure Control

Consider the following example: Alice participates in a social network in which users can interact with each other to share information. Even if shared information is never globally visible (which can, for instance, partially be achieved by the privacy settings in Facebook [79]) the interaction between users can allow information shared by Alice to quickly spread throughout the network: assume, for instance, that Alice initially shares some piece of information with a small group of her friends in the social network. Through various gossiping mechanisms, these friends can then share this piece of information with other people, and these people in turn with further people, causing Alice to very quickly lose control over the piece of information she initially shared with only a limited number of her friends.

Similar behavior can today be observed with information that *goes viral*. Even if the information was always visible globally, it is the sharing/re-tweeting/liking of the shared information and the associated transitive diffusion of it that causes the information to finally reach a massive audience, much larger than often was intended. On the other hand, many entities also seek to use this mechanism for their own benefits, i.e. in the instance of viral marketing [27, 89].

2.3.2 Exposure Control in Closed Settings

In traditional, closed settings, we solve this issue by enforcing access control mechanisms using a central control mechanism (for instance the file system on a computer)

and thereby providing guarantees of non-interference. While there have been some approaches towards adopting access control mechanisms to open settings [7], these have the drawback of requiring a trusted third party. This requirement makes such systems often very hard to be adopted in practice.

However, even if a trusted third party were to be established, the open nature in which users can interact with shared data makes controlling its exposure impossible in open settings: by simply generating a copy of the data in question (e.g. by making a screenshot of a picture or writing down a one to one copy of a text document), a user can easily step outside of the controlled ecosystem and will be free to further share the data. As such, even encryption based mechanisms were we only allow a specific set of initial users to receive our data by, e.g., encrypting it with their public keys will not protect the data from being shared further.

In Chapter 6, we develop an alternative approach to controlling the exposure of information in decentralized, open settings: we enable the user to share information in a controlled manner in order to minimize the expected exposure of the information by making an a priori exposure estimation/minimization based on the structure of the network and the set of notes the user wants to share information with. Effectively, however, we cannot provide provable guarantees with this approach, but instead resort to estimating the risk of a privacy violation and try to enable the user making informed decisions about their information sharing behavior.

2.4 From Privacy Guarantees to Privacy Risk Assessment

A common issue that we observe by going from the closed settings to the decentralized and open settings is that providing provable privacy guarantees seems increasingly difficult. Since we lack the central entity that controls access to information, enforcing privacy mechanisms in a way that still preserves utility for the users seems impossible: on the one hand, other users in the network (or adversaries) have direct access to any information that is shared, instead of being limited to how they access the information. Achieving similar privacy guarantees as with Differential Privacy would then require to add a significant amount of noise that would destroy any utility of the information that is shared. On the other hand, once information has left control of the original user, access to their information can longer be regulated since the communication network is open and decentralized.

However, we think that, even if provable privacy guarantees are not attainable, providing the user with sound privacy risk assessment can meaningfully enhance their privacy. Throughout this dissertation we will therefore follow the approach of providing meaningful privacy risk assessments to the user, ideally enabling them to make informed decisions about the information dissemination behavior.

3
Related Work

We next give an overview of work related to the results presented in the following chapters. This includes work on traditional privacy notions and information disclosure in social networks relevant to Chapters 4 and 5, as well as related work from the subject area of diffusion networks for chapters 6 and 7.

3.1 Privacy in Closed-world Settings

The notion of privacy has been exhaustively discussed for specific settings such as statistical databases, as well as for more general settings. Since we already discussed the notions of k-anonymity [103], l-diversity [74] t-closeness [70] and Differential Privacy [30] in Section 2.2.3 in great detail, we will now discuss further such notions.

A major point of criticism of Differential Privacy, but also the other existing privacy notions, found in the literature [59] is the (often unclear) trade-off between utility and privacy that is incurred by applying database sanitation techniques to achieve privacy. Several works have shown that protection against attribute disclosure cannot be provided in settings that consider an adversary with arbitrary auxiliary information [25, 31, 32]. We later show, as sanity check, that in our formalization of privacy in open settings, general non-disclosure guarantees are indeed impossible to achieve. By providing the necessary formal groundwork in this paper, we hope to stimulate research on *assessing* privacy risks in open settings, against explicitly spelled-out adversary models.

Kasiviswanathan and Smith [55] define the notion of ϵ-semantic privacy to capture general non-disclosure guarantees. We define our adversary model in a similar fashion as in their formalization and we use ϵ-semantic privacy to show that general non-disclosure guarantees cannot be meaningfully provided in open settings.

Several extensions of the above privacy notions have been proposed in the literature to provide privacy guarantees in use cases that differ from traditional database privacy [8, 14, 117, 48, 119, 15]. These works aim at suitably transforming different settings into a database-like setting that can be analyzed using differential privacy. Such a transformation, however, often abstracts away from essential components of these settings, and as a result achieve impractical privacy guarantees. As explained in Section 2.2.3, the open web setting is particularly ill-suited for such transformations.

Specifically for the use case in Online Social Networks (in short, OSNs), many works [72, 119, 15] apply the existing database privacy notions for reasoning about attribute disclosure in OSNs. These works generally impose a specific structure on OSN data, such as a social graph, and reason about the disclosure of private attributes through this structure. Liu and Terzi [72], and Zhou et al. [119] adopt the notions of k-anonymity and l-diversity to protect nodes in social graph data. Here, [72] build anonymity sets by node degrees and achieve anonymity by adding and deleting edges in the social graph, whereas [119] considers the while neighborhood of a node for its anonymity subset, and discusses the complexity of finding private node-subsets based on this criterium. Both approaches, however, suffer from the same problems these techniques have in traditional statistical data disclosure, where an adversary with auxiliary information can easily infer information about any specific user.

There also exist several works that apply Differential Privacy to achieve privacy in social graphs [92, 88, 109, 111, 15]. Sala et al. [92] and Xaio et al. [111] both propose

differentially private sanitization mechanism that perturb the original social graph to achieve differential privacy with respect to the edges in the graph. Chen et al. [15] further extend these differential privacy approaches to the case where several nodes in the network are correlated, and thus traditional differential privacy would fail to protect the social graph.

Proserpio et al. [88], and Wang and Wu [109], on the other hand, propose graph synthesis mechanisms that measure key structural properties of a given social graph, and then generate a new, random social graph that closely resembles the original graph in these structural properties. All of these approaches, however, remain static, and it is assumed that the data can be globally sanitized in order to provide protection against information disclosure. Again, as discussed in Section 2.2.3, this does not apply to the open web setting with its highly unstructured dissemination of data.

3.2 Data Processing in OSNs

There has been a significant amount of work in processing and understanding data obtained from OSNs [117, 62, 48, 75, 18, 21, 84, 93, 11, 91, 24, 35, 51, 52, 22]. Different Natural Language Processing techniques have successfully been used to understand user-generated text content and derive information from it. This includes inferring location information [18], political alignment [21], health information [84, 93] and other attributes [11, 75, 35, 22, 51, 52] for specific users, but also detecting events and incidents that affect many different users [91, 24]. Even simply evaluating the context in which a user submits a search query can lead to attribute disclosure [54]. This exemplifies again that OSNs specifically, and the Internet in general, provide an ubiquitous source of data that can provide a significant amount of information about users. Zhaleva et al. [117] show that mixed public and private profiles do not necessarily protect the private part of a profile since they can be inferred from the public part. Heatherly et al. [48] similarly show how machine learning techniques can be used to infer private information from publicly available information due to the often existing correlation between what one might consider harmless information and sensitive information one seeks to hide or protect. Kosinksi et al. [62] moreover show that machine learning techniques can indeed be used to predict personality traits of users and their online behavior.

Several works show that stylometric features of text can be leveraged to identify the author of a given text [60, 4, 1]. They consider attributes such as n-grams frequencies, usage of punctuation, etc., to match authors to texts. Inspired by these works, we follow a simplified approach of utilizing unigram frequencies as attributes of our statistical models for the experimental evaluation of our privacy model in chapters 4 and 5.

Scerri et al. present the digital.me framework [95, 94] which attempts to unify a user's social sphere across different OSNs by, e.g., matching the profiles of the same user across these OSNs. While their approach is limited to the closed environment they consider, their work provides interesting insights into identity disclosure in more open settings.

3.3 Matching Identities

A number of works propose profile matching schemes that leverage profile attributes provided by users themselves such as their names, locations or bios [39, 85, 2, 83] to match profiles of the same user across different social networks. Of particular interest is the study by Goga et al. [39], which shows that it is possible to accurately link 30% of Twitter identities to their matching identities on Facebook. However, it is not possible to exactly pinpoint the matching identity for the remaining 70% of Twitter identities. This insight serves as perfect motivation for our paper: can we build a framework that assesses the individual risk of a user to have his identities matched across sites.

Other studies matched identities by exploiting friends lists or the graph structure of social networks [114, 64, 61, 81]. For example, Narayanan et al. [81] showed the feasibility to de-anonymize the friendship graph of a social network on a large scale using the friendship graph of another social network as auxiliary information.

For geo-location data specifically, Cecaj et al [13] investigate the possibility of matching identities in call detail records to identities in social networks. They characterize the uniqueness of an identity by the number of data points required to uniquely identify an identity and then try to match this uniquely identified identity to its social network profiles using statistical methods similar to the ones proposed in this paper.

Finally, other papers proposed schemes to match identities by exploiting the content generated by users [38, 77, 50]. For example, Mishari et al. [77] show that domain reviews could also be linked across different sites by exploiting the language model of the authors.

Several other works show that even stylometric features of text can be leveraged to identify the author of a given text [4]. Inspired by these works, we also use language models to represent identities. Note that our risk assessment framework can work with any kind of attributes, but for this study we limit ourselves at using language models as attributes.

In Chapter 5, we will see that anonymity within one social network alone is often not enough to protect against linkability, since the open nature of the online social network setting does not allow for the same equivalence of anonymity and unlinkability that we typically see in closed settings such as statistical databases.

3.4 Privacy in Online Social Networks

As discussed above, there is a growing body of research that utilizes commonly used machine learning and information retrieval techniques to extract critical information from user profiles and the content that users disseminate in OSNs. Only few works have tried to develop protection mechanisms against such methods. Most among them (e.g., [63, 76]) have focused on the protection of so-called Personally Identifiable Information (PII) introduced in privacy and data-protection legislation [26], which constitute a fixed set of entity attributes that even in isolation supposedly lead to the unique identification of entities. Narayanan and Shmatikov, however, show that the differentiation between key attributes that identify entities, and sensitive attributes that need to be protected, is not appropriate for privacy in pervasive online settings such as the Inter-

net [81, 80]. Technical methods for identifying and matching entities do not rely on the socially perceived sensitivity of attributes for matching, but rather any combination of attributes can lead to successful correlation of corresponding profiles. Our privacy model treats every type of entity attribute as equally important for privacy and allows for the identification of context-dependent, sensitive attributes.

3.5 Statistical Language Models

In chapters 4 and 5, we will use statistical language models to represent the information contained in the text published by user profiles. Statistical Language Models for information retrieval have first been introduced by Ponte and Croft [87] as an alternative approach for document retrieval and are inspired by language models for Speech Recognition and Natural Language Processing [100, 90]. They have subsequently been focus of a long line of research (examples include [65, 116, 105, 115]) that further develop the basic statistical language model approach and its benefits. While Statistical Language Models have not been shown to perform better than other established retrieval methods [115], we found that the Statistical Language Model formulation is closer than other options to what we require in expressing and solving indistinguishability problems that arise in computer security. Approach does allow for structural insight into the retrieval problem by structuring documents by their content: our goal is to use this structural insight for reasoning about indistinguishability in security problems.

Hiemstra et al. [49] introduce the notion of parsimonious language models that extract information that is specific to a document when compared to a background collection of documents. In the context of privacy analysis, this allows us to identify critical properties of entities that distinguish them from other entities in the same collection. From this information, one can then formulate countermeasures that hide these critical properties, and thereby increase the privacy of this entity.

3.6 Controlling Information Propagation in Social Networks

In chapters 6 and 7, we consider the problem of controlling the propagation of shared information in a social network due to user interacting with each other (e.g. gossiping). On a basic level, there already exist approaches and implementations for controlling access to shared information in the literature. A basic such approach to controlling the visibility of information has been implemented in Facebook through social access control lists [79]. They, however, only take into account the direction transmission of information from the information source to other nodes in the target. In particular, social access control lists do not consider the transitive diffusion of information throughout the network, which we consider in this work. In follow up work, Mondal et al. [78] propose the general notion of exposure for controlling the diffusion of shared information instead. The definition of exposure that we use in Chapters 6 and 7 is heavily motivated by their informal definitions.

Some recent works investigate the flow of sensitive information in popular social network despite privacy control mechanism designed to protect such sensitive information (usually due to utility requirements by the provider) [73, 71]. They thus show

issues on the implementation level where the expected behavior of privacy policies does not coincide with their real behavior.

Several works investigate the prediction of privacy settings for social networks [99, 112]. They show that user privacy preferences are dynamic and context dependent, but at the same time can often be predicted with fairly high accuracy. The approach presented in this paper requires the user to define propagation policies for each of their actions separately. An automated system that accurately predicts these propagation policies can therefore be helpful to substantially increase the usability of any practical system that adopts our approach.

3.7 Diffusion Networks

In this work, we leverage diffusion networks to study the propagation of information throughout a social network. Diffusion Networks have extensively been used in the past to model and predict the propagation of various quantities throughout connected systems, such as social and behavioral influence [47, 96], epidemiology [107] and viral marketing [27, 89]. Kempe et al. [57] unify some of the previous definition in two different propagation models, name the independent cascade model and the linear threshold model. In the independent cascade model, each infected node i has a single chance to infect each of its neighbors j with a infection likelihood $p_{i,j}$ that is a model specific parameters.

In the linear threshold model, on the other hand, each node i has a randomly chosen infection threshold θ_i and a weight $w_{i,j}$ to each neighbor j. A node is then infect if the sum of the weights of all infected neighbors exceeds the nodes infection threshold.

In our work, we leverage the continuous time diffusion networks proposed by Gomez-Rodriguez et al. [43] to model the propagation of information through social networks. These continuous time diffusion networks are an extension of the regular independent cascade model: instead of having a single chance of infecting a neighboring node, each edge is now associated with a continuous infection likelihood function f. The likelihood of infection at time t is then given by $f(t)$. Continuous time models allow us to take into account the time dimension when formulating privacy policies, making it possible to much easier satisfy privacy constraints when, e.g., they are only required to hold for a very short time. We will introduce the continuous time diffusion network in more detail in Chapter 6.

3.8 Diffusion Model Inference

The algorithms presented in chapters 6 and 7 assume a diffusion network that accurately represents the diffusion behavior of the real (social) networks that we want to work on. The corresponding *model inference* problem of finding such an accurate representation is an actively researched topic in the literature. Gomez-Rodriguez et al. show that the model inference problem for continuous-time diffusion networks is NP-hard, but allows for a constant factor approximation due to the submodularity of the corresponding objective function [41]. In [42], Gomez-Rodriguez et al. present an approach for inferring the diffusion parameters for dynamic networks. While the current

implementation of XpoMin relies on static networks, investigating options for adopting to dynamically changing networks (by, for instance, continuously generating distance network samples based on ever changing diffusion network and using a sliding windows approach) seems to be a promising direction for future work.

3.9 Influence in Diffusion Networks

One of the main uses of diffusion models is to estimate the *influence* of a node within a diffusion network. In the instance of information propagation, influence represents the expected number of nodes in the network that will receive a piece of information shared by a given node. Identifying most influential nodes in a diffusion network can help deciding with whom to share a piece of information, or which parts of the network are particularly critical.

The algorithmic problem of computing the influence of a given node has been shown to be #P-hard for the linear threshold model by Chen et al. [17] as well as for the independent cascade model by Wang et al. [108]. Later, Gomez-Rodriguez et al. [46] present the same complexity for estimating influence in continuous-time diffusion networks since they constitute a generalization of the independent cascade model.

Since influence estimation is a major building block for further algorithms on diffusion networks (influence maximization and minimization), several approximation algorithms have been developed to deal with the algorithmic complexity of the influence estimation problem. Wang et al. [108] and Gomez-Rodriguez et al. [46] both propose Monte-Carlo simulation based approaches to estimate the influence of a node in independent cascade and continuous-time diffusion networks.

Du et al. [28] furthermore present a purely simulation-based, scalable influence estimation algorithm for continuous-time diffusion networks: they simply simulate the information spreading to obtain samples for the propagation time of information between two nodes and then average the sampled times over a large number of samples to get an accurate estimate for the average number of infected nodes. In our work in chapter 7, we adapt this simulation based algorithm for the estimation of *exposure*, which is a generalization of influence.

3.10 Influence Minimization

Our main goal in chapters 6 and 7 will be to minimize the exposure of information in a network by carefully choosing the nodes with which we initially share a piece of information. For regular influence, there already exists some work on the topic of influence minimization in diffusion networks. In contrast to the work presented in this work, however, existing approaches seek to find ideal modifications to the network that minimize the influence with the goal of minimizing the reach of undesirable qualities within the diffusion network.

Khalil et al., for instance, consider the issue of optimizing the network structure itself to minimize influence [58]. They find edges and nodes in the network that, when removed, cause a maximal reduction of influence of a selected number of seed nodes that probabilistically infect the network with information. In our work, our goal is

not to modify the network, but to optimally satisfy propagation policies within a fixed network.

Yao et al., on the other hand, propose an alternative mechanism in which a number of nodes in the network are blocked from interacting with the rest of the network in order to minimize the spread of a piece of information in the network [113].

3.11 Heterogenous Diffusion Models

Barbieri et al. propose a topic aware extension of traditional diffusion models [10]. In this model, the transmission likelihood does not only depend on a pairwise transmission rate, but also on the topic dependent adaptation rate specific to each node. Our basic approach of using propagation policies to control information propagation could easily be adapted to such a model. However, the implications for the algorithmic tractability of the corresponding optimization problems are unclear.

Du et al. present an approach to model diffusion networks with heterogeneous transmission functions [29]. In contrast to homogeneous transmission functions that we use in this work and which are agnostic to the type of information that is shared, heterogeneous transmission functions allow the modeling of diffusion processes that also depend on the type of information that is shared. Adopting such heterogeneous models might allow for more useful exposure minimizations where we take into account the propensity of a node to share a specific type of information when deciding with which nodes to share a piece of information of that type.

4
(k,d)-anonymity

Anonymity in Open Settings

4.1 Motivation

The Internet has undergone dramatic changes in the last two decades, evolving from a mere communication network to a global multimedia platform in which billions of users not only actively exchange information, but increasingly conduct sizable parts of their daily lives. While this transformation has brought tremendous benefits to society, it has also created new threats to online privacy that existing technology is failing to keep pace with. Users tend to reveal personal information without considering the widespread, easy accessibility, potential linkage and permanent nature of online data. Many cases reported in the press show the resulting information disclosure risks, which range from public embarrassment and loss of prospective opportunities (e.g., when applying for jobs or insurance), to personal safety and property risks (e.g., when sexual offenders or burglars learn users' whereabouts online). The resulting privacy awareness and privacy concerns of Internet users have been further amplified by the advent of the Big-Data paradigm and the aligned business models of personalized tracking and monetizing personal information in an unprecedented manner.

4.2 Problem Description

Prior research on privacy has traditionally focused on closed database settings – characterized by a complete view on structured data and a clear distinction of key- and sensitive attributes – and has aimed for strong privacy guarantees using global data sanitization. These approaches, however, are inherently inadequate if such closed settings are replaced by open settings as described above, where unstructured and heterogeneous data is being disseminated, where individuals have a partial view of the available information, and where global data sanitization is impossible and hence strong guarantees have to be replaced by probabilistic privacy assessments.

As of now, *even the basic methodology is missing* for offering users technical means to comprehensively assess the privacy risks incurred by their data dissemination, and their daily online activities in general. Existing privacy models such as k-anonymity [103], l-diversity [74], t-closeness [70] and the currently most popular notion of Differential Privacy [30] follow a database-centric approach that is inadequate to meet the requirements outlined above. We refer the reader to Section 2.2.3 for further discussions on existing privacy models.

4.3 Contributions

In this paper, we present a rigorous methodology for quantitatively assessing identity disclosure risks in open settings. Concretely, the paper makes the following three tangible contributions: (1) a formal framework for reasoning about the disclosure of personal information in open settings, (2) an instantiation of the framework for reasoning about the identity disclosure problem, and (3) an evaluation of the framework on a collection of 15 million comments collected from the Online Social Network Reddit.

A Formal Framework for Privacy in Open Settings. We propose a novel framework

for addressing the essential challenges of privacy in open settings, such as providing a data model that is suited for dealing with unstructured dissemination of heterogeneous information through various different sources and a flexible definition of user-specific privacy requirements that allow for the specification of context-dependent privacy goals. In contrast to most existing approaches, our framework strives to assess the degree of exposure individuals face, in contrast to trying to enforce an individual's privacy requirements. Moreover, our framework technically does not differentiate between non-sensitive and sensitive attributes a-priori, but rather starts from the assumption that all data is equally important and can lead to privacy risks. More specifically, our model captures the fact that the sensitivity of attributes is highly user- and context-dependent by deriving information sensitivity from each user's privacy requirements. As a sanity check we prove that hard non-disclosure guarantees cannot be provided for the open setting in general, providing incentive for novel approaches for assessing privacy risks in the open settings.

Reasoning about Identity Disclosure in Open Settings. We then instantiate our general privacy framework for the specific use case of identity disclosure. Our framework defines and assesses identity disclosure (i.e., identifiability and linkability of identities) by utilizing entity similarity, i.e., an entity is private in a collection of entities if it is sufficiently similar to its peers. At the technical core of our model is the new notion of d-convergence, which quantifies the similarity of entities within a larger group of entities. It hence provides the formal grounds to assess the ability of any single entity to blend into the crowd, i.e., to hide amongst peers. The d-convergence model is furthermore capable of assessing identity disclosure risks specifically for single entities. To this end, we extend the notion of d-convergence to the novel notion of (k, d)-anonymity, which allows for entity-centric identity disclosure risk assessments by requiring d-convergence in the local neighborhood of a given entity. Intuitively, this new notion provides a generalization of k-anonymity that is not bound to matching identities based on pre-defined key-identifiers.

Empirical Evaluation on Reddit. Third, we perform an instantiation of our identity disclosure model for the important use case of analyzing user-generated text content in order to characterize specific user profiles. We use unigram frequencies extracted from user-generated content as user attributes, and we subsequently demonstrate that the resulting unigram model can indeed be used for quantifying the degree of anonymity of – and ultimately, for differentiating – individual entities. For the sake of exposition, we apply this unigram model to a collection of 15 million comments collected from the Online Social Network Reddit. The computations were performed on two Dell PowerEdge R820 with 64 virtual cores each at 2.60GHz over the course of six weeks. Our evaluation shows that (k, d)-anonymity suitably assesses an identity's anonymity and provides deeper insights into the data set's structure.

4.4 A Framework for Privacy in Open Settings

In this section, we first develop a user model that is suited for dealing with the information dissemination behavior commonly observed on the Internet and the associated

identity disclosure risks. We then formalize our adversary model and show, as a sanity check, that hard privacy guarantees against identity disclosure cannot be achieved in open settings. We conclude by defining privacy goals for the problem of in open settings through user-specified privacy requirements from which we then derive a new definition of information sensitivity suited to the problem of identity disclosure in open settings.

4.4.1 Modeling Information in Open Settings

We first define the notion of entity models and restricted entity models. These models capture the behavior of these entities and in particular describe which attributes an entity exhibits publicly.

Definition 1 (Entity Model)**.** *Let $\mathcal{A}$ be the set of all attributes. The* entity model θ_ϵ *of an entity ϵ provides for all attributes $\alpha \in \mathcal{A}$ an* attribute value $\theta_\epsilon(\alpha) \in \mathsf{dom}(\alpha) \cup \{\mathsf{NULL}\}$ *where $\mathsf{dom}(\alpha)$ is the domain over which the attribute α_i is defined.*

The domain $\mathsf{dom}(\theta)$ of an entity model θ is the set of all attributes $\alpha \in \mathcal{A}$ with value $\theta(\alpha) \neq \mathsf{NULL}$.

An entity model thus corresponds to the information an entity can publicly disseminate. With the specific null value NULL we can also capture those cases where the entity does not have any value for that specific attribute.

In case the adversary has access to the full entity model, a set of entity models basically corresponds to a database with each attribute $\alpha \in \mathcal{A}$ as its columns. In the open setting, however, an entity typically does not disseminate all attribute values, but instead only a small part of them. We capture this with the notion of restricted entity models.

Definition 2 (Restricted Entity Model)**.** *The* restricted entity model $\theta_\epsilon^{\mathcal{A}'}$ *is the entity model of ϵ restricted to the non empty attribute set $\mathcal{A}' \neq \emptyset$, i.e.,*

$$\theta_\epsilon^{\mathcal{A}'}(\alpha) = \begin{cases} \theta_\epsilon(\alpha), & \text{if } \alpha \in \mathcal{A}' \\ \mathsf{NULL}, & \text{otherwise} \end{cases}$$

In the online setting, each of the entities above corresponds to an online profile. A user u usually uses more than one online service, each with different profiles $P_1^u, \ldots, P_l^u$. We thus define a user model as the collection of the entity models describing each of these profiles.

Definition 3 (User Model / Profile Model)**.** *The* user model $\theta_u = \{\theta_{P_1^u,}, \ldots, \theta_{P_1^u}\}$ *of a user u is a set of the entity models $\theta_{P_1^u,}, \ldots, \theta_{P_1^u,}$, which we also call* profile models.

With a user model that separates the information disseminated under different profiles, we will be able to formulate privacy requirements for each of these profiles separately. We will investigate this in Section 4.4.4.

4.4.2 Adversary Model

In the following we formalize the adversary we consider for privacy in open settings. In our formalization, we follow the definitions of a semantic, Bayesian adversary introduced by Kasiviswanathan and Smith [55].

For any profile P, we are interested in what the adversary Adv learns about P observing publicly available information from P. We formalize this learning process through *beliefs* on the models of each profile.

Definition 4 (Belief)**.** *Let $\mathcal{P}$ be the set of all profiles and let $\mathcal{D}_{\mathcal{A}}$ be the set of all distributions over profile models. A* belief $b = \{b_P | P \in \mathcal{P}\}$ *is a set of distributions $b_P \in \mathcal{D}_{\mathcal{A}}$.*

We can now define our privacy adversary in open settings using the notion of belief above.

Definition 5 (Adversary)**.** *An adversary* Adv *is a pair of* prior *belief b and world knowledge κ, i.e.,* $\mathsf{Adv} = (b, \kappa)$.

The adversary Adv's prior belief b represents his belief in each profile's profile model before makes any observations. This prior belief can, in particular, also include background knowledge about each profile P. The world knowledge κ of the adversary represents a set of inference rules that allow him to infer additional attribute values about each profile from his observations.

We next define the publicly observations based on which the adversary learns additional information about each profile.

Definition 6 (Publication Function)**.** *A* publication function *G is a randomized function that maps each profile model θ_P to a restricted profile model $G(\theta_P) = \theta_P^{\mathcal{A}'}$ such that there is at least one attribute $\alpha \in \mathcal{A}'$ with $\theta_P(\alpha) = G(\theta_P)(\alpha)$.*

The publication function G reflects which attributes are disseminated publicly by the user through his profile P. G can, in particular, also include local sanitization where some attribute values are perturbed. However, we do require that at least one attribute value remains correct to capture utility requirements faced in open settings.

A public observation now is the collection of all restricted profile models generated by a publication function.

Definition 7 (Public Observation)**.** *Let $\mathcal{P}$ be the set of all profiles, and let G be a publication function. The* public observation *$\mathcal{O}$ is the set of all restricted profile models generated by G, i.e., $\mathcal{O} = \{G(\theta_P) | P \in \mathcal{P}\}$.*

The public observation $\mathcal{O}$ essentially captures all publicly disseminated attribute values that can be observed by the adversary. Given such an observation $\mathcal{O}$, we can now determine what the adversary Adv learns about each profile by determining his *a-posteriori belief.*

Definition 8 (A-Posteriori Belief)**.** *Let $\mathcal{P}$ be the set of all profiles. Given an adversary* $\mathsf{Adv} = (b, \kappa)$ *and a public observation $\mathcal{O}$, the adversary's* a-posteriori belief $\bar{b} = \{\bar{b}_P \in \mathcal{D}_{\mathcal{A}} | P \in \mathcal{P}\}$ *is determined by applying the Bayesian inference rule, i.e.,*

$$\bar{b}_P[\theta | \mathcal{O}, \kappa] = \frac{Pr[\mathcal{O} | \kappa, \theta] \cdot b_P[\theta]}{\sum_{\theta'} Pr[\mathcal{O} | \kappa, \theta'] \cdot b_P[\theta']}.$$

Here, the conditional probability $Pr[\mathcal{O}|\kappa, \theta]$ describes the likelihood that the observational $\mathcal{O}$ is created by the specific entity model θ.

We will utilize the a-posteriori belief of the adversary to reason about the violation of the user specified privacy requirements in Section 4.4.4.

4.4.3 Inapplicability of Statistical Privacy Notions

In the following, we formally show that traditional non-disclosure guarantees, e.g., in the style of Differential Privacy, are not possible in open settings.

Kasiviswanathan and Smith [55] provide a general definition of non-disclosure they call ϵ-privacy. In their definition, they compare the adversary Adv's a-posteriori beliefs after observing the transcript t generated from a database sanitazitaion mechanism $\mathcal{F}$ applied on two adjacent databases with n rows: first on the database x, leading to the belief $\overline{b}_0[.|t]$, and secondly on the database x_{-i}, where a value in the ith row in x is replaced by a default value, leading to the belief $\overline{b}_i[.|t]$.

Definition 9 (ϵ-semantic Privacy [55])**.** *Let $\epsilon \in [0, 1]$. A randomized algorithm $\mathcal{F}$ is ϵ-semantically private if for all belief distributions b on D^n, for all possible transcripts, and for all $i = 1 \ldots n$:*

$$\mathsf{SD}(\overline{b}_0[.|t], \overline{b}_i[.|t]) \leq \epsilon.$$

Here, SD is the total variation distance of two probability distributions.

Definition 10. *Let X and Y be two probability distributions over the sample space D. The* total variation distance SD *of X and Y is*

$$\mathsf{SD}(X, Y) = \mathsf{max}_{S \subset D} \left[Pr[X \in S] - Pr[Y \in S] \right].$$

Kasiviswanathan and Smith [55] show that ϵ-differential privacy is essentially equivalent to ϵ-semantic privacy.

In our formalization of privacy in open settings, varying a single database entry corresponds to changing the value of a single attribute α in the profile model θ_P of a profile P to a default value. We denote this modified entity model with θ_P^α, and the thereby produced a-posteriori belief by $\overline{b}_P^\alpha$. A profile P would then be ϵ-semantically private if for any modified profile model θ_P^α, the a-posteriori belief of adversary Adv does not change by more than ϵ.

Definition 11 (ϵ-semantic Privacy in Open Settings)**.** *Let $\epsilon \in [0, 1]$. A profile P is ϵ-semantically private in open settings if for any attribute α,*

$$\mathsf{SD}(\overline{b}_P[.|\mathcal{O}], \overline{b}_P^\alpha[.|\mathcal{O}]) \leq \epsilon$$

where $\overline{b}_P$ and $\overline{b}_P^\alpha$ are the a-posteriori beliefs of the adversary after observing the public output of θ_P and θ_P^α respectively.

As expected, we can show that ϵ-semantic privacy can only hold for $\epsilon = 1$ in open settings.

Theorem 1. *For any profile model θ_P and any attribute α, there is an adversary* Adv *such that*

$$\mathsf{SD}(\overline{b}[.|\mathcal{O}], \overline{b}^{\alpha}[.|\mathcal{O}]) \geq 1.$$

Proof. Let Adv have a uniform prior belief, i.e., all possible profile models have the same probability, and empty world knowledge κ. Let α be the one attribute that remains the same after applying the publication function G. Let x be the original value of this attribute α and let x^* be the default value that replaces x.

Observing the restricted profile model $\theta_P[\mathcal{A}']$ without any additional world knowledge will lead to an a-posteriori belief, where the probability of the entity model θ with $\theta[\mathcal{A}'] = \theta_P[\mathcal{A}']$ and NULL everywhere else, is set to 1.

Conversely, the modified setting will result in an a-posteriori belief that sets the probability for the entity model θ^* to one, where θ^* is constructed for the modified setting as θ above. Thus $\overline{b}[\theta|\mathcal{O}] = 1$, whereas $\overline{b}^{\alpha}[\theta|\mathcal{O}] = 0$, and hence $\mathsf{SD}(\overline{b}[.|\mathcal{O}], \overline{b}^{\alpha}[.|\mathcal{O}]) = 1$. □ □

Intuitively, the adversary can easily distinguish differing profile models because a) he can directly observe the profiles publicly available information, b) he chooses which attributes he considers for his inference and c) only restricted, local sanitization is available to the profile. Since these are elementary properties of privacy in open settings, we can conclude that hard security guarantees in the style of differential privacy are impossible to achieve in open settings.

However, we can provide an assessment of the disclosure risks by explicitly fixing the a-priori knowledge and the attribute set considered by the adversary. While we no longer all-quantify over all possible adversaries, and therefore lose the full generality of traditional non-disclosure guarantees, we might still provide meaningful privacy assessments in practice. We further discuss this approach in Section 4.4.5, and follow this approach in our instantiation of the general model for assessing the likelihood of identity disclosure in Section 4.5.

4.4.4 User-Specified Privacy Requirements

In the following we introduce user-specified privacy requirements that allow us to formulate privacy goals against identity disclosure that are user- and context-dependent. These can then lead to restricted privacy assessments instead of general privacy guarantees that we have shown to be impossible in open setting in the previous section.

As pointed out in [81, 80], in practice, there are no personal attributes that are inherently more sensitive than other attributes. In fact, the sensitivity of any personal attribute depends on the context of use.

While there is prior work that aims to infer the sensitivity of information from the context of interactions in online social networks, this work is empirical and only evaluates which type of information can potentially be sensitive, and not whether and how the same user can deem different information sensitive in different contexts. We instead assume user-specified privacy policies, i.e., policies that allow each user to explicitly specify which attributes should not appear in the public user model and hence

should be hidden from the adversary. For example, the user might have an online social network profile solely for the purpose of discussing political topics, and additionally maintains a separate personal profile that is kept detached from the political profile. Privacy requirements would then be that the personal profile does not leak any information about the user's political opinions, whereas the political account does not leak any personal information.

Definition 12 (Privacy Policy)**.** *A* privacy policy $\mathcal{R}$ *is a set of privacy requirements* $r = (P, \{\alpha_i = x_i\})$ *which require that profile* P *should never expose the attribute values* x_i *for the attributes* $\alpha_i \in \mathcal{A}$*.*

By setting privacy requirements in a per-profile basis we capture an important property of information dissemination in open settings: users utilize different profiles for different context (e.g., different online services) assuming these profiles remain separate and specific information is only disseminated under specific circumstances.

Given the definition of privacy policies, we now define the violation of a policy by considering the adversary's a-posterior belief $\bar{b}$, as introduced in Section 4.4.2.

Definition 13 (Privacy Policy Satisfaction / Violation)**.** *Let* $\mathsf{Adv} = (b, \kappa)$ *be an adversary with a-posteriori belief* $\bar{b}$*, and let* $\theta[\alpha = x]$ *be the set of all entity models that have the value* x *for the attribute* α*. A profile* P_i^u σ-satisfies *a user's privacy requirement* $r_j^u = (P, \{\alpha_i = x_i\})$*, written* $P_i^u \models_\sigma r_j^u$*, if*

- $P = P_i^u$
- $\forall \alpha_i : \sum_{\theta \in \theta[\alpha_i = x_i]} \bar{b}_P[\theta|\mathcal{O}, \kappa] \leq \sigma$

and σ-violates *the user's privacy requirement otherwise.*

A user model θ_u σ-satisfies *a user* u*'s privacy policy* $\mathcal{R}_u$*, written* $\theta_u \models_\sigma \mathcal{R}_u$*, if all profile models* $\theta_{P_i^u}$ σ*-satisfy their corresponding privacy requirements, and* σ-violates *the privacy policy otherwise.*

The above attributes can also take the form of "P belongs to the same user as P'", effectively restricting which profiles should be linked to each other. We will investigate this profile linkability problem specifically in Section 4.5.

4.4.5 Sensitive Information

In contrast to the closed-world setting, with its predefined set of sensitive attributes that automatically defines the privacy requirements, a suitable definition of information sensitivity w.r.t. identity disclosure in open settings is still missing. In the following, we derive the notion of sensitive information from the user privacy requirements we defined in Section 4.4.4.

Definition 14 (Sensitive Attributes)**.** *A set of attributes* $\mathcal{A}^*$ *is* sensitive *for a user* u *in the context of her profile* P_i^u *if* u*'s privacy policy* $\mathcal{R}_u$ *contains a privacy requirement* $r = (P_i^u, \mathcal{A}' = X)$ *where* $\mathcal{A}^* \subseteq \mathcal{A}'$*.*

Here, we use the notation $\mathcal{A} = X$ as vector representation for $\forall \alpha_i \in \mathcal{A} : \alpha_i = x_i$.

Sensitive attributes, as defined above, are not the only type of attributes that are worth to protect: In practice, an adversary can additionally infer sensitive attributes from other attributes through statistical inference using a-priori knowledge. We call such attributes that allow for the inference of sensitive attributes *critical attributes*.

Definition 15 (Critical Attributes)**.** *Given a set of attributes $\mathcal{A}^*$, let P be a profile with $\mathsf{dom}(\theta_P) \supseteq \mathcal{A}$, and let P' be the profile with the restricted profile model $\theta_{P'} = \theta_P^{\mathcal{A}'}$, where $\mathcal{A}' = \mathsf{dom}(\theta_P) \setminus \mathcal{A}^*$.*

The set of attributes $\mathcal{A}^*$ is σ-critical *for the user u that owns the profile P and an adversary with prior belief b_P and world knowledge κ, if u's privacy policy $\mathcal{R}_u$ contains a privacy requirement r such that P σ-violates r but P' does not.*

Critical information require the same amount of protection as sensitive information, the difference however being that critical information is only protected for the sake of protecting sensitive information.

As a direct consequence of the definition above, sensitive attributes are also critical.

Corollary 1. *Let $\mathcal{A}$ be a set of sensitive attributes. Then $\mathcal{A}$ is also* 0*-critical.*

Another consequence we can draw is that privacy requirements will always be satisfied if no critical attributes are disseminated.

Corollary 2. *Let $\mathcal{O}$ be a public observations that does not include any critical attributes for a user u and an adversary* Adv*. Then u's privacy policy $\mathcal{R}_u$ is σ-satisfied against* Adv*.*

The corollary above implies that, while we cannot provide general non-disclosure guarantees in open settings, we can provide privacy assessments for specific privacy requirements, given an accurate estimate of the adversary's prior beliefs.

While privacy risk assessments alone are not satisfactory from a computer security perspective, where we usually require hard security guarantees quantified over all possible adversaries, the fact remains that we are faced with privacy issues in open settings that are to this day unanswered for due to the impossibility of hard guarantees in such settings. Pragmatically thinking, we are convinced that we should move from impossible hard guarantees to more practical privacy assessments instead. This makes particularly sense in settings where users are not victims of targeted attacks, but instead fear attribute disclosure to data-collecting third parties.

4.5 Anonymity in Open Settings

In the following we instantiate the general privacy model introduced in the last section to reason about the likelihood that two profiles of the same user are linked by the adversary in open settings. We introduce the novel notion of (k, d)-anonymity with which we assess anonymity and linkability based on the similarity of profiles within an online community.

To simplify the notation we introduce in this section, we will, in the following, talk about matching *entities* ϵ and ϵ' the adversary wants to link, instead of profiles P_1 and P_2 that belong to the same user u. All definitions introduced in the general framework above naturally carry over to entities as well.

4.5.1 Model Instantiation for Linkability

In the linkability problem, we are interested in assessing the likelihood that two matching entities ϵ and ϵ' can be linked, potentially across different online platforms. The corresponding privacy requirements, as introduced in Section 4.4.4, are $r_1 = (\epsilon, \alpha_L)$ and $r_2 = (\epsilon', \alpha_L)$, where α_L is the attribute that ϵ and ϵ' belong to the same user. Consequently, we say that these entities are unlinkable if they satisfy the aforementioned privacy requirements.

Definition 16 (Unlinkability)**.** *Two entities ϵ and ϵ' are σ-unlinkable if*

$$\{\theta_\epsilon, \theta_{\epsilon'}\} \models_\sigma \{r_1, r_2\}.$$

4.5.2 Anonymity

To assess the identity disclosure risk of an entity ϵ within a collection of entities $\mathcal{E}$, we use the following intuition: ϵ is anonymous in $\mathcal{E}$ if there is a subset $\mathcal{E}' \subseteq \mathcal{E}$ to which ϵ is very similar. The collection $\mathcal{E}$' then is an anonymous subset of $\mathcal{E}$ for ϵ.

To assess the similarity of entities within a collection of entities, we will use a distance measure dist on the entity models of these entities. We will require that this measure provides all properties of a metric.

A collection of entities in which the distance of all entities to ϵ is small (i.e., $\leq$ a constant d) is called d-convergent for ϵ.

Definition 17. *A* collection *of entities $\mathcal{E}$ is d-convergent for ϵ if $\mathsf{dist}(\theta_\epsilon, \theta_{\epsilon'}) \leq d$ for all $\epsilon' \in \mathcal{E}$.*

Convergence measures the similarity of a collection of individuals. Anonymity is achieved if an entity can find a collection of entities that are all similar to this entity. This leads us to the definition of (k, d)-anonymity, which requires a subset of similar entities of size k.

Definition 18. *An entity ϵ is* (k, d)-anonymous *in a collection of entities $\mathcal{E}$ if there exists a subset of entities $\mathcal{E}' \subseteq \mathcal{E}$ with the properties that $\epsilon \in \mathcal{E}$, that $|\mathcal{E}'| \geq k$ and that $\mathcal{E}'$ is d-convergent.*

An important feature of this anonymity definition is that it provides anonymity guarantees that can be derived from a subset of all available data, but continue to hold once we consider a larger part of the dataset.

Corollary 3. *If an entity is (k, d)-anonymous in a collection of entities $\mathcal{E}$, then it is also (k, d)-anonymous in the collection of entities $\mathcal{E}' \supset \mathcal{E}$.*

Intuitively, (k, d)-anonymity is a generalization of the classical notions of k-anonymity to open settings without pre-defined quasi-identifiers. We schematically illustrate such anonymous subsets in Figure 4.1.

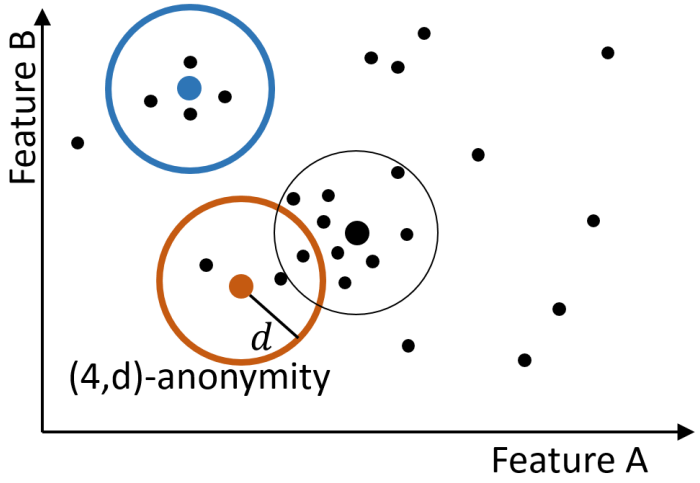


Figure 4.1: Anonymity in crowdsourcing systems.

4.5.3 Entity Matching

We define the notion of *matching* identities. As before, we use the distance measure dist to assess the similarity of two entities.

Definition 19. *An entity ϵ c-matches an entity ϵ' if* $\mathsf{dist}(\theta_\epsilon, \theta_{\epsilon'}) \leq c$.

Similarly, we can also define the notion of one entity matching a collection of entities.

Definition 20. *A collection of entities $\mathcal{E}$* c-matches *an entity ϵ' if all entities $\epsilon \in \mathcal{E}$ c-match ϵ'.*

Assuming the adversary only has access to the similarity of entities, the best he can do is comparing the distance of all entities $\epsilon \in \mathcal{E}$ to ϵ' and make a probabilistic choice proportional to their relative distance values.

Now, if the matching identity ϵ^* is d-convergent in $\mathcal{E}$ the, all entities in $\mathcal{E}$ will have a comparatively similar distance to ϵ'.

Lemma 1. *Let $\mathcal{E}$ be d-convergent for ϵ^*. If ϵ^* c-matches ϵ', then $\mathcal{E}$ $(c+d)$-matches ϵ'.*

Proof. Since $\mathcal{E}$ is d-convergent for ϵ^*, $\forall \epsilon' \in \mathcal{E}$: $\mathsf{dist}(\epsilon^*, \epsilon') \leq d$. Using the triangle inequality, and the fact that ϵ^* c-matches the entity ϵ', we can bound the distance of all entities $\epsilon \in \mathcal{E}$ to ϵ' by $\forall \epsilon'' \in \mathcal{E}$: $\mathsf{dist}(\epsilon, \epsilon') \leq c + d$. Hence $\mathcal{E}$ $(c+d)$-matches the entity ϵ'. □

Hence, the matching entity ϵ^* does not c-match ϵ' for a small value of c, the adversary Adv he will have a number of possibly matching entities that are similarly likely to match ϵ'.

We get the same result if not the whole collection $\mathcal{E}$ is convergent, but if there exists a subset of convergent entities that allows the target to remain anonymous.

Corollary 4. *Let ϵ' be (k,d)-anonymous in $\mathcal{E}$. If ϵ' c-matches an entity ϵ then there is a subset $\mathcal{E}' \subseteq \mathcal{E}$ of size at least k which $(c+d)$-matches ϵ.*

4.5.4 Identity Disclosure

We now provide an assessment for the likelihood that an adversary successfully links matching entities. The notion of (k, d)-anonymity we introduced in Chapter 4 informally means that an entity is able to hide among at least k other entities with a similarity of at least d. This ideally implies that an adversary is not able, or at least is severely hindered, in uniquely identifying an entity in a anonymous subset. In the following we bound the probability with which an adversary can successfully identify the unique entity ϵ from a collection of entities $\mathcal{E} = \{\epsilon_1, \ldots, \epsilon_n\}$ that corresponds to an outside target entity ϵ'.

We assume that the adversary uses the similarity of the candidate entities to his target entity ϵ' to make his decision. The likelihood that the adversary chooses a specific entity ϵ^* then is the relative magnitude of $\mathsf{dist}(\epsilon^*, \epsilon)$, i.e.

$$Pr[\mathsf{Adv} \text{ chooses } \epsilon^*] = 1 - \frac{\mathsf{dist}(\epsilon^*, \epsilon')}{\sum_{\epsilon \in \mathcal{E}} \mathsf{dist}(\epsilon, \epsilon')}.$$

We can now bound the likelihood with which a specific entity ϵ^* would be chosen by the adversary if ϵ^* is (k, d)-anonymous.

Theorem 2. *Let the matching entity ϵ^* of the entity ϵ' in the collection $\mathcal{E} = \{\epsilon_1, \ldots, \epsilon_n\}$ be (k, d)-anonymous in $\mathcal{E}$. Furthermore let ϵ^* c-match ϵ'. Then an adversary $\mathsf{Adv} = (b, \emptyset)$ with uniform prior belief b and with empty world knowledge that only observes the similarity of entities links the entity ϵ^* to ϵ' with a likelihood of at most $t \leq 1 - \frac{c}{c+(k-1)(c+d)}$.*

Proof. Let $\mathcal{E}^*$ be the (k, d) anonymous subset of ϵ^* in $\mathcal{E}$. Let t^* be the likelihood of identifying ϵ^* from $\mathcal{E}^*$. Then clearly $t < t^*$ since we remove all possible, but wrong candidates in $\mathcal{E} \setminus \mathcal{E}^*$.

Since ϵ^* c-matches ϵ', by Lemma 1, we can upper bound the distance of each entity in $\mathcal{E}^*$ to ϵ', i.e.,

$$\forall \epsilon \in \mathcal{E}^* : \mathsf{dist}(\epsilon, \epsilon') \leq c + d$$

We can now bound t^* as follows:

$$\begin{aligned} t^* &= Pr[\mathsf{Adv} \text{ chooses } \epsilon] \\ &= 1 - \frac{c}{c + (k-1)(\sum\limits_{\epsilon \in \mathcal{E}^* \setminus \{\epsilon^*\}} \mathsf{dist}(\epsilon, \epsilon'))} \leq 1 - \frac{c}{c + (k-1)(c+d)} \end{aligned}$$

□

Theorem 2 shows that, as long as entities remain anonymous in a suitably large anonymous subset of a collection of entities, an adversary will have difficulty identifying them with high likelihood. Recalling our unlinkability definition from the beginning of the section, this result also implies that ϵ^* is σ-unlinkable for $\sigma = t$.

Corollary 5. *Let the matching entity ϵ^* of the entity ϵ' in the collection $\mathcal{E} = \{\epsilon_1, \ldots, \epsilon_n\}$ be (k, d)-anonymous in $\mathcal{E}$. Then ϵ^* and ϵ' are σ-unlinkable for $\sigma = 1 - \frac{c}{c+(k-1)(c+d)}$ against an adversary $\mathsf{Adv} = (b, \emptyset)$ with uniform prior belief and empty world knowledge that only observes entity similarity.*

In Section 4.6.6 we present experiments that evaluate the anonymity and linkability of individuals in the Online Social Network Reddit, and measure how well they can be identified from among their peers.

4.5.5 Limitations

The quality of the assessment provided by the d-convergence model largely depends on the adversarial prior belief: in our results above, we assume an adversary without any prior knowledge. In practice, however, the adversary might have access to prior beliefs that can help him in his decision making. Therefore, turning such assessments into meaningful estimates in practice requires a careful estimation of prior knowledge by, e.g., producing a more accurate profile model: the problem of comprehensive profile building for entities in an open setting is an open question that has been examined somewhat in the literature [12, 23, 16, 98, 9], but on the whole still leaves a lot of space for future work.

This concludes the formal definitions of our d-convergence model. In the next sections, we instantiate it for identity disclosure risk analyses based on user-generated text-content and apply this instantiation to the OSN Reddit.

4.6 Anonymity Evaluation on Reddit

While the main focus of this paper is to present the actual privacy model as such, the following experiments are meant to provide first insights into the application of our framework, without taking overly complex adversarial capabilities into account. The evaluation can easily be extended to a more refined model of an adversary without conceptual difficulties.

We first articulate the goals of this evaluation, and then, secondly, describe the data collection process, followed by defining the instantiation of the general framework we use for our evaluation in the third step. Fourth, we introduce the necessary processing steps on our dataset, before we finally discuss the results of our evaluation.

4.6.1 Goals

In our evaluation, we aim at validating our model by conducting two basic experiments. First, we want to empirically show that, our model instantiation yields a suitable abstraction of real users for reasoning about their privacy. To this end, profiles of the same user should be more similar to each other (less distant) than profiles from different users.

Second, we want to empirically show that a larger anonymous subset makes it more difficult for an adversary to correctly link the profile. Thereby, we inspect whether anonymous subsets provide a practical estimate of a profile's anonymity.

Given profiles with anonymous subsets of similar size, we determine the percentage of profiles which the adversary can match within the top k results, i.e., given a source profile, the adversary computes the top k most similar (less distant) profiles in the

other subreddit. We denote this percentage by *precision@k* and correlate it to the size of the anonymous subsets.

We fix the convergence of the anonymous subsets to be equal to the matching distance between two corresponding profiles. Our intuition is that, this way, the anonymous subset captures most of the profiles an adversary could potentially consider matching.

4.6.2 Data-Collection

For the empirical evaluation of our privacy model, we use the online social network Reddit [104] that was founded in 2005 and constitutes one of the largest discussion and information sharing platforms in use today. On Reddit, users share and discuss topics in a vast array of topical subreddits that collect all topics belonging to one general area; e.g. there are subreddits for world news, tv series, sports, food, gaming and many others. Each subreddit contains so-called submissions, i.e., user-generated content that can be commented on by other users.

To have a ground truth for our evaluation, we require profiles of the same user across different OSNs to be linked. Fortunately, Reddit's structure provides an inherent mechanism to deal with this requirement. Instead of considering Reddit as a single OSN, we treat each subreddit as its own OSN. Since users are identified through the same pseudonym in all of those subreddits, they remain linkable across the subreddits' boundaries. Hence our analysis has the required ground truth. The adversary we simulate, however, is only provided with the information available in the context of each subreddit and thus can only try to match profiles across subreddits. Ground truth in the end allows us to verify the correctness of his match.

To build up our dataset, we built a crawler using Reddit's API to collect comments. Recall that subreddits contain submissions that, in turn, are commented by the users. For our crawler, we focused on the large amount of comments because they contain a lot of text and thus are best suitable for computing the unigram models.

Our crawler operates in two steps that are repeatedly executed over time. During the whole crawling process, it maintains a list of already processed users. In the first step, our crawler collects a list of the overall newest comments on Reddit from Reddit's API and inserts these comments into our dataset. In the second step, for each author of these comments who has not been processed yet, the crawler also collects and inserts her latest $1,000$ comments into our dataset. Then, it updates the list of processed users. The number of $1,000$ comments per user, is a restriction of Reddit's API.

In total, during the whole September 2014, we collected more than 40 million comments from over $44,000$ subreddits. The comments were written by about $81,000$ different users which results in more than 2.75 million different profiles.

The whole dataset is stored in an anonymized form in a MySQL database and is available upon request.

4.6.3 Ethical Concerns

For our evaluation, we only collected publicly available, user-generated text content from the social media system Reddit and replaced the pseudonyms under which this

content was posted with randomized, numerical identifiers. In our evaluation, we did not infer any further information about the users; in particular we did not directly link any profiles, but used the pseudonym information to match the same user's content across different subreddits. We thus do not infer any further sensitive information (through linking) than what is already publicly made available by each user on the Reddit platform.

Since our institutes do not have an IRB, we consulted the opinion of a local privacy lawyer, who confirmed that our research is in accordance with the Max Planck Society's ethics guidelines as well as with the applicable German data protection legislation (§28 BDSG) at that time.

4.6.4 Model Instantiation

On Reddit, users only interact with each other by by posting comments to text of link submissions. Reddit therefore does not allow us to exploit features found in other social networks, such as friend links or other static data about each user. On the other hand, this provides us with the opportunity to evaluate the linkability model introduced in Section 4.5 based dynamic, user-generated content, in this case user-generated text content.

Since we only consider text content, we instantiate the general model from the previous sections with an unigram model, where each attribute is a word unigram, an its value is the frequency with which the unigram appears in the profiles comments. Such unigram models have succesfully been used in the past to characterize the information within text content and to correlate users across different online platforms [38, 77].

Definition 21 (Unigram Model)**.** *Let $\mathcal{V}$ be a finite vocabulary. The* unigram model *$\theta_P = p_i$ of a profile is a set of* frequencies *$p_i \in [0, \ldots, 1]$ with which each unigram $w_i \in \mathcal{V}$ appears in the profile P. Each frequency p_i is determined by*

$$p_i = \frac{\mathsf{count}(w_i, P)}{\sum_{w \in \mathcal{V}} \mathsf{count}(w, P)}$$

Since the unigram model essentially constitutes a probability distribution, we instantiate our distance metric dist with the Jensen-Shannon divergence [33]. The Jensen-Shannon divergence is a symmetric extension of the Kullback-Leiber divergence has been shown to be successful in many related information retrieval scenarios.

Definition 22. *Let P and Q be two statistical models over a discrete space Ω. The Jensen-Shannon divergence is defined by*

$$D_{\mathsf{JS}} = \frac{1}{2} D_{\mathsf{KL}}(P||M) + \frac{1}{2} D_{\mathsf{KL}}(Q||M)$$

where D_{KL} is the Kullback-Leibler divergence

$$D_{\mathsf{KL}}(P||Q) = \sum_{\omega \in \Omega} log\left(\frac{P(\omega)}{Q(\omega)}\right) P(\omega)$$

and M is the averaged distribution $M = \frac{1}{2}(P + Q)$.

In the following, we will use the square-root of the Jensen-Shannon divergence, constituting a metric, as our distance measure, i.e., $\mathsf{dist} = \sqrt{D_{\mathsf{JS}}}$.

4.6.5 Data-Processing

The evaluation on our dataset is divided into sequentially performed computation steps, which include the normalization of all comments, the computation of unigram models for each profile, a filtering of our dataset to keep the evaluation tractable, the computation of profile distances and the computation of (k, d)-anonymous subsets.

Normalizing Comments. Unstructured, heterogeneous data, as in our case, may contain a variety of valuable information about a user's behavior, e.g., including formatting and punctuation. Although we could transform these into attributes, we do not consider them here for the sake of simplicity.

In order to get a clean representation to apply the unigram model on, we apply various normalization steps, including transformation to lower case, the removal of Reddit formatting and punctuation except for smilies. Moreover, we apply a encoding specific normalization, replace URLs by their hostnames and shorten repeated characters in words like `cooool` to a maximum of three. Finally, we also filter out a list of 597 stopwords from the comments. Therefore, we perform six different preprocessing steps on the data, which we describe in more detail in the following.

1. **Convert to lower case letters:** In our statistical language models, we do not want to differentiate between capitalized and lowercased occurrences of words. Therefore, we convert the whole comment into lower case.

2. **Remove Reddit formatting:** Reddit allows users to use a wide range of formatting modifiers that we divide into two basic categories: formatting modifiers that influence the typography and the layout of the comment, and formatting modifiers that include external resources into a comment. The first kind of modifier, named layout modifiers, is stripped off the comment, while leaving the plain text. The second kind of modifier, called embedding modifiers, is removed from the comment completely.

   One example for a layout modifier is the asterisk: When placing an asterisk both in front and behind some text, e.g., `*text*`, this text will be displayed in italics, e.g., *text*. Our implementation removes these enclosing asterisks, because they are not valuable for computing statistical language models for n-grams and only affect the layout. Similarly, we also remove other layout modifiers such as table layouts, list layouts and URL formatting in a way that only the important information remains.

   A simple example for embedding modifiers are inline code blocks: Users can embed arbitrary code snippets into their comments using the ` modifier. Since these code blocks do not belong to the natural language part of the comment and only embed a kind of external resource, we remove them completely. In addition to code blocks, the category of embedding modifiers also includes quotes of other comments.

3. **Remove stacked diacritics:** In our dataset, we have seen that diacritics are often misused. Since Reddit uses Unicode as its character encoding, users can

create their own characters by arbitrarily stacking diacritics on top of them. To avoid this kind of unwanted characters, we first normalize the comment by utilizing the unicode character composition, which tries to combine each letter and its diacritics into a single precombined character. Secondly, we remove all remaining diacritic symbols from the comment. While this process preserves most of the normal use of diacritics, it is able to remove all unwanted diacritics.

4. **Replace URLs by their hostname:** Generally, a URL is very specific and a user often does not include the exact same URL in different comments. However, it is much more common that a user includes different URLs that all belong to the same hostname, e.g., `www.mypage.com`. Since our statistical language models should represent the expected behavior of a user in terms of used words (including URLs), we restrict all URLs to their hostnames.

5. **Remove punctuation:** Most of the punctuation belongs to the sentence structure and, thus, should not a part of our statistical language models. Therefore, we remove all punctuation except for the punctuation inside URLs and smilies. We do not remove the smilies, because people are using them in a similar role as words to enrich their sentences: Every person has her own subset of smilies that she typically uses. To keep the smilies in the comment, we maintain a list of 153 different smilies that will not be removed from the comment.

6. **Remove duplicated characters:** In the internet, people often duplicate characters in a word to add emotional nuances to their writing, e.g., `cooooooooool`. But sometimes the number of reduplicated characters varies, even if the same emotion should be expressed. Thus, we reduce the number of duplicated characters to a maximum of 3, e.g., `coool`. In practice, this truncation allows us to differentiate between the standard use of a word and the emotional variation of it, while it does not depend on the actual number of duplicated characters.

Computing Unigram Models. From the normalized data, we compute the unigram frequencies for each comment. Recall that our dataset consists of many subreddits that each form their own OSN. Thus, we aggregate the corresponding unigram frequencies per profile, per subreddit, and for Reddit as a whole. Using this data, we compute the word unigram frequencies for each comment as described in Section 4.6.4.

Since a subreddit collects submissions and comments to a single topic, we expect the unigrams to reflect its topic specific language. Indeed, the 20 most frequently used unigrams of a subreddit demonstrate that the language adapts to the topic. As an example, we show the top 20 unigrams (excluding stopwords) of Reddit and two sample subreddits *Lost* and *TipOfMyTongue* in Table 4.1. As expected, there are subreddit specific unigrams that occur more often in the context of one subreddit than in the context of any other subreddit. For example, the subreddit *Lost* deals with a TV series that is about the survivors of a plane crash and its aftermath on an island. Unsurprisingly, the word *island* is the top unigram in this subreddit. In contrast, the subreddit *TipOfMyTongue* deals with the failure to remember a word from memory and, thus, has the word *remember* in the list of its top three unigrams.

Top	Reddit		subreddit: Lost		subreddit: TipOfMyTongue	
	Unigram	Frequency	Unigram	Frequency	Unigram	Frequency
1.	people	4,127,820	island	832	www.youtube.com	3663
2.	time	2,814,841	show	750	song	1,542
3.	good	2,710,665	lost	653	remember	1,261
4.	gt	2,444,240	time	580	en.wikipedia.org	1,100
5.	game	1,958,850	people	527	sounds	1,007
6.	pretty	1,422,640	locke	494	solved	924
7.	2	1,413,118	season	431	movie	918
8.	lot	1,385,167	jacob	429	find	829
9.	work	1,352,292	mib	372	:)	786
10.	1	1,184,029	jack	310	game	725
11.	3	1,124,503	episode	280	time	678
12.	great	1,070,299	ben	255	thinking	633
13.	point	1,063,239	good	250	good	633
14.	play	1,060,985	monster	237	www.imdb.com	584
15.	years	1,032,270	lot	220	video	583
16.	bad	1,008,607	gt	182	pretty	570
17.	day	989,180	character	165	youtu.be	569
18.	love	988,567	walt	163	mark	548
19.	find	987,171	man	162	edit	540
20.	shit	976,928	dharma	162	post	519

Table 4.1: Top 20 unigrams of Reddit and two sample subreddits Lost and TipOfMyTongue.

Filtering the Dataset. To reduce the required amount of computations we restrict ourselves to *interesting profiles*. We define an interesting profile as one that contains at least 100 comments and that belongs to a subreddit with at least 100 profiles. Additionally, we dropped the three largest subreddits from our dataset to speed up the computation.

In conclusion, this filtering results in $58,091$ different profiles that belong to $37,935$ different users in $1,930$ different subreddits.

Distances Within and Across Subreddits. Next, we compute the pairwise distance within and across subreddits using our model instantiation. Excluding the distance of profiles to themselves, the minimal, maximal and average distance of two profiles within subreddits in our dataset are approximately 0.12, 1 and 0.79 respectively. Across subreddits, the minimal, maximal and average distance of two profiles are approximately 0.1, 1 and 0.85 respectively.

Anonymous Subsets. Utilizing the distances within subreddits, we can determine the anonymous subsets for each profile in a subreddit. More precisely, we compute the anonymous subset for each pair of profiles from the same user. We set the convergence

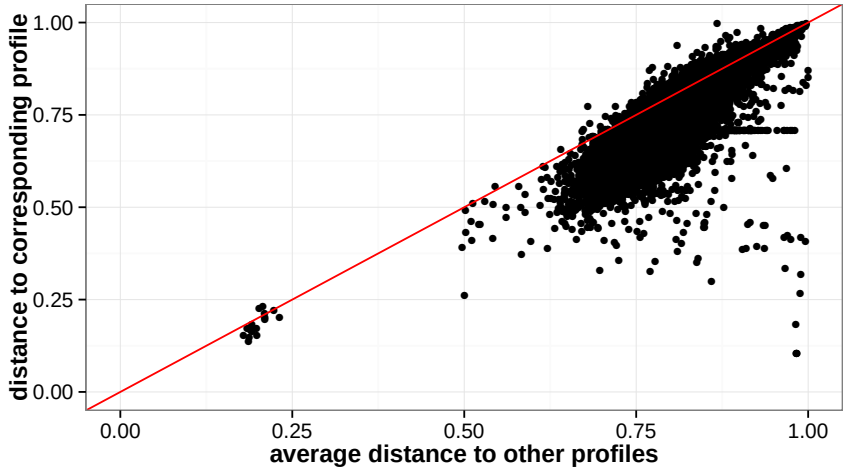


Figure 4.2: The average distance between a profile in subreddit s and all profiles in s' versus the matching distance between the profile and its correspondence in s'.

d to the matching distance between both profiles and determine the size of the resulting anonymous subset.

4.6.6 Evaluation and Discussion

In this subsection, we inspect and interpret the results of our experiments with regard to our aforementioned goals. Therefore, we first start by giving evidence that our approach indeed provides a suitable abstraction of real users for reasoning about their privacy.

To this end, we compare the distance of matching profiles to the average distance of non-matching profiles. In particular, for each pair of profiles from the same user in subreddits s and s', we plot the average distance from the profile in s to the non-matching profiles in s' in relation to the distance to the matching profile in s' in Figure 4.2. The red line denotes the function $y = x$ and divides the figure into two parts: if a point lies below the line through the origin, the corresponding profiles match better than the average of the remaining profiles. Since the vast majority of datapoints is located below the line, we can conclude that profiles of the same user match better than profiles of different users.

Our second goal aimed at showing that anonymous subsets indeed can be used to reason about the users' privacy. Therefore, we investigate the chances of an adversary to find a profile of the same user within the top k matches and relate its chance to the size of the profile's anonymous subset. More precisely, given multiple target profiles with similar anonymous subset sizes, we determine the, so called, precision@k, i.e., the ratio of target profiles that occur in the top k ranked matches (by ascending distance from the source profiles). We relate this precision@k to the anonymous subset sizes with a convergence d set to the distance between the source and target profiles, and we group the anonymous subset sizes in bins of size 10.

In our evaluation, we considered $k \in \{1, 5, 10, 20\}$, which all yield very similar

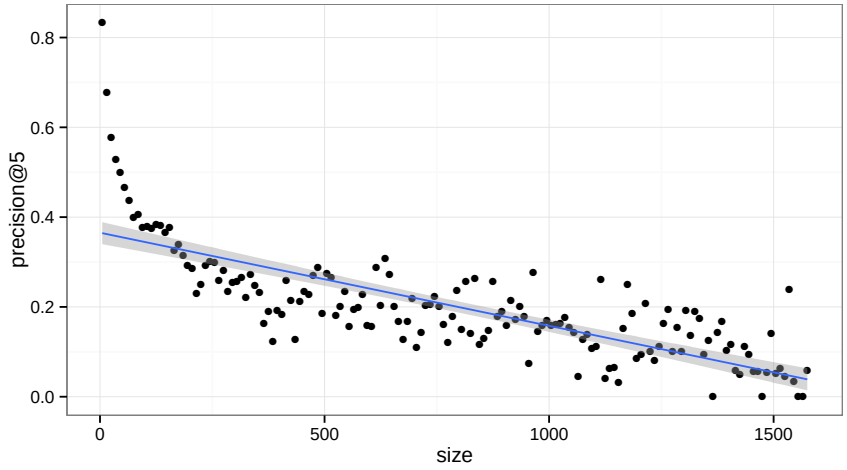


Figure 4.3: The anonymous subset size correlated to the precision an adversary has if considering the top 5 profiles as matching.

results. Exemplarily, we correlate the aforementioned measures for $k = 5$ in Figure 4.3, clearly showing that an increasing anonymous subset size correlates with an increasing uncertainty – i.e., decreasing precision – for the adversary.

4.7 Conclusion and Future Work

We presented a user-centric privacy framework for reasoning about privacy in open web settings. In our formalization, we address the essential challenges of protecting against identity disclosure in open settings: we defined a comprehensive data model that can deal with the unstructured dissemination of heterogeneous information, and we derived the sensitivity of information from user-specified and context-sensitive privacy requirements. We showed that, in this formalization of privacy in open settings, hard security guarantees in the sense of Differential Privacy are impossible to achieve. We then instantiated the general framework to reason about the identity disclosure problem. The technical core of our identity disclosure model is the new notion of (k, d)-anonymity that assesses the anonymity of entities based on their similarity to other entities within the same community. We applied this instantiation to a dataset of 15 million user-generated text entries collected from the Online Social Network Reddit and showed that our framework is suited for the assessment of anonymity in Online Social Networks: with increasing anonymous subset size k, the likelihood of a successful linking attack decreases.

As far as future work is concerned, many directions are highly promising. First, our general framework only provides a static view on privacy in open settings. Information dissemination on the Internet, however, is, in particular, characterized by its highly dynamic nature. Extending the model presented in this paper with a suitable transition system to capture user actions might lead to powerful system for monitoring privacy risks in dynamically changing, open settings. Second, information presented in Online

Social Networks is often highly time-sensitive, e.g., shared information is often only valid for a certain period of time, and personal facts can change over time. Explicitly including timing information in our entity model will hence further increase the accuracy of the entity models derived from empirical evidence. Finally, our privacy model is well-suited for the evaluation of protection mechanisms for very specific privacy requirements, and new such mechanisms with provable guarantees against restricted adversaries can be developed. On the long run, we pursue the vision of providing the formal foundations for comprehensive, trustworthy privacy assessments and, ultimately, for developing user-friendly privacy assessment tools.

5

Profile Linkability Despite Anonymity in Social Media Systems

5.1 Motivation

Social media systems, where any user can join the system and contribute content, are becoming widely popular. Examples of social media systems include blogging sites like Twitter and LiveJournal, social bookmarking sites like Delicious and Reddit, and peer-opinion sites like Yelp, Amazon, and eBay reviews. To enable users to contribute freely and without fear, these sites need to offer their users *anonymity*. Today, many systems allow users to operate using pseudonymous identities that can be created without providing any certification by trusted authorities and where users determine what information they choose to reveal about themselves. For instance, many Twitter users do not provide (or deliberately provide fake) information about their real names, bios, or profile photos when creating identities.

Many users participate in different social media sites assuming different pseudonymous identities under the belief that their identities across different sites cannot be linked. However, recently researchers have shown that adversaries can exploit seemingly innocuous and latent information such as location patterns [38] and linguistic patterns in *public posts* [77] to link even pseudonymous identities that a user has created across different sites. Such attempts to aggregate and link user data across multiple social media sites in order to reveal a more comprehensive profile of the information sources have many commercial applications [101], but they also raise serious privacy concerns for the users of these sites.

5.2 Problem Description

In this chapter, we examine the degree to which the anonymity of a user's identity can be used to estimate the linkability threats that are *inherent* to the *publicly visible content* contributed by a user to social media sites. That is, we evaluate linkability threats assuming that the only data that is available for linking a user's identities are the contents of the public posts written using the identities. In practice, an adversary might have additional data about a user (e.g., non-public data such as a user's IP address or a user's real name) that might help them link the user's identities. However, we consider only public posts of the user as (i) they are available to all adversaries and (ii) they represent the minimum amount of information a user reveals by participating in the social media site. Consequently, we consider the unavoidable linkability threat that arises from a user's content contributions to different social media sites.

Our work is motivated by the relation of linkability and anonymity of a user's identities in a traditional database setting. In such a setting, anonymity usually requires equality within an anonymity set, which naturally implies unlinkability of the user's identities. The same, however, cannot directly be applied to the linkability of user posts in social media systems like Facebook, Twitter or Reddit since, on such platforms, information is presented in a highly unstructured manner: traditional privacy models, such as k-anonymity [103], l-diversity [74], t-closeness [70], or differential privacy [31], have been defined over well-structured databases and cannot be applied to user posts (e.g., it is not clear what the quasi-identifiers and sensitive attributes in this context are). Moreover, it is unclear how differentially private noising would work on natural

language posts.

5.3 Contributions

We leverage the notion of d-convergence introduced in the previous chapter that extends the notion of k-anonymity over structured data sets to unstructured data sets: For a user identity u in a social media system, (k, d)-anonymity captures the largest k subset of identities containing u such that every identity within the subset is within a divergence (or dissimilarity) threshold of d from u.

Using (k, d)-anonymity, we evaluate whether anonymity in one social media system allows us to estimate the risk of linkability threats across social media systems. Specifically, we address the following two questions: (i) Can the knowledge of (k, d)-anonymity of users in an online social media system be used to estimate their *relative linkability risks*, i.e., estimate whether one user is more at risk of her identities being linked than other users? (ii) To what extent does combining knowledge of (k, d)-anonymity of a user in a social media system with information about their matching identity in a different social media system improve the linkability assessment?

We use an extensive data set of over 15 million comments posted by users across $1,930$ topical communities in the Reddit social media system. Using potential strategies of a rational adversary, we analyze the correlations between the (k, d)-anonymity of a user's identity and the estimated risk of the identity being matched to determine the utility of the (k, d)-anonymity measure.

Our findings yield several valuable insights about the relation between anonymity and linkability. First, the ranking of identities by the size of their (k, d)-anonymity set positively correlates with the matching set size (i.e., the number of identities the adversary considers as potentially matching). This is what we have also observed in the experimental evaluation in chapter 4. However, this correlation is fairly weak and we thus conclude that *anonymity alone is not sufficient to assess linkability risks on social media systems.* Second, we find that enriching the anonymity sets found by (k, d)-anonymity with information about the matching identities yields linkability risk assessment that is much more useful in practice. Using the local matching set μ that we derive by combining anonymity sets and information about matching identities we can succesfully estimate the size of the matching set: in over 74% of the cases, the size of the local matching set μ is at least 0.8 times the actual matching set size of the adversary.

Outline We begin by introducing required background knowledge and motivating our work in Section 5.4. We then develop the relative and absolute linkability measures in Section 5.5. In Section 5.6, we introduce the Reddit data set we use for our evaluations. Using this data set, we then evaluate, in Section 5.7, both linkability measures and show that anonymity alone is not a good measure of linkability, but extending anonymity with information about matching identities can provide a good measure of linkability. We finally conclude in Section 5.8.

5.4 Background and Motivation

Before examining how well anonymity can be used to assess linkability threats that allow an adversary to link identities across sites, we first have to discuss the terminology we use in the remainder of the paper and provide the background on key concepts that underlie our work.

5.4.1 Domains and Identities

The term *identity* denotes the profile created by a user in a social media system. A *domain* is the collection of identities within a social media system. A pair of identities within different domains is called *matching* if they belong to the same user.

5.4.2 Identity Representation and Similarity

The first challenge in addressing the anonymity and linkability threats in social media systems is to find a suitable representation of identities. Given that the *only* information that we presume to know about an identity are its public posts, we represent each identity by fitting a statistical model to the identity's textual posts.

The simplest way to construct such a statistical model would be to determine the relative frequency of each word unigram used by an identity. Specifically, we represent identities through the same unigram-statistical language model that we have also used in the previous chapter: it captures the relative frequency with which the identity uses a specific unigram w: i.e., given a vocabulary $\mathcal{V}$ of word unigrams and the collection of comments $C_{\mathcal{I}}$ by $\mathcal{I}$, the identity model $\theta_{\mathcal{I}}$ is defined by

$$Pr[w \mid \theta_{\mathcal{I}}] = \frac{\mathsf{count}(w, C_{\mathcal{I}})}{\sum_{w' \in \mathcal{V}} \mathsf{count}(w', C_{\mathcal{I}})}.$$

While this identity model is fairly simple, it is sufficient to assess the relation between anonymity and linkability in social media systems that allow the sharing of user-generated text content. In Section 5.7, we investigate various more complex models. The general observations, however, stay the same. It would also be possible to incorporate other sources of information – as for example pictures, videos or location – into the identity model. Naturally, the precise anonymity and linkability risk of an identity will then change with such an extended model that includes a wider variety of features, however, in this paper we are rather interested in gaining conceptual clarity into the ways anonymity and linkability relate to each other, rather than estimating the precise linkability risks of an identity in a specific system and under specific scenarios.

To measure how *similar* two identities are we again use the Jensen-Shannon divergence [33] D_{JS}.[1] As discussed in the previous chapter, the Jensen-Shannon divergence is a symmetric variant of the popular Kullback-Leibler divergence, which has been used with large success to determine the similarity of probability distributions (and therefore statistical models), and the square root of D_{JS} provides a full-fledged metric. In the

[1] We also tested other metrics such as Cosine similarity in Section 5.7.1, but the results were not affected significantly.

remainder of the paper we will talk about the *distance* $\mathsf{dist}(\mathcal{I}, \mathcal{I}') = \sqrt{D_{\mathsf{JS}}(\theta_{\mathcal{I}}, \theta_{\mathcal{I}'})}$ of identities, induced by this divergence measure, instead of their similarity, to provide a better, intuitive understanding.

5.4.3 Adversarial Matching Strategy

In this paper, we consider an adversary that tries to link a source identity $\mathcal{I}_{\mathcal{S}}$ within a source domain $\mathcal{S}$ to the matching target identity $\mathcal{I}_{\mathcal{T}}$ within a target domain $\mathcal{T}$. We assume that the adversary has at her disposal (i) the posts of all identities in both domains and (ii) a small ground-truth set of matching identities across both domains. These are standard assumptions made by the majority of previous work in this area [39].

The matching process of the adversary Adv consists of four steps: (i) Adv computes the pairwise similarity between all identities in $\mathcal{S}$ and all identities in $\mathcal{T}$; (ii) he computes the likelihood of any two identities to belong to the same user based on their similarity[2]; (iii) he then ranks all pairs of identities according to their likelihood of belonging to the same user; and (iv) the adversary chooses a threshold th on the likelihood measure (according to how accurately he wants to link identities) and links all the identities that are above the threshold. The threshold choice is the standard trade-off between recall (i.e., the fraction of identities linked out of all matching identities) and precision (i.e., the probability that the identities linked are actually matching identities) calculated over the ground-truth set of matching identities. This strategy is consistent with the strategy employed by the majority of previous works on matching identities. We already discussed several such works in Chapter 3.

While the matching strategy we consider in this paper corresponds to a rational adversary, who wants to increase the number of identities he can link correctly, this adversary model does not necessarily represent the worst case adversary; and an adversary could simply choose to not be rational. As pointed out by Backes et al. [6] it is, in general, impossible to provide unlinkability guarantees against arbitrary adversaries in open and unstructured settings that we consider in this work.

5.4.4 Linkability of Identities

Through his choice of the threshold value th (see Section 5.4.3) the adversary defines the set of identities within the target domain $\mathcal{T}$ that he considers *potentially matching* the source identity $\mathcal{I}_{\mathcal{S}}$: we call this set the *matching set* $\mathcal{M}(th)$ of the adversary for identity $\mathcal{I}_{\mathcal{S}}$. We illustrate such a matching set in Figure 5.1. The matching set is the set of identities from which the adversary cannot sufficiently distinguish which target identity $\mathcal{I}_{\mathcal{T}}$ (cf. Figure 5.1) is related to $\mathcal{I}_{\mathcal{S}}$.

We can therefore quantify the linkability of a user's identities using this matching set: the bigger $\mathcal{M}(th)$ is, the less likely it is that the adversary will link $\mathcal{I}_{\mathcal{S}}$ to $\mathcal{I}_{\mathcal{T}}$. Note that the size of the matching set of an identity depends on the threshold th chosen by the adversary. In this paper, we will consider both scenarios where we know and where

[2] An adversary can consider the similarity between two identities ($\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{T}}$) as the likelihood of them to belong to the same user, or he can compute more complex functions that, in addition to the similarity between $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{T}}$, take into account the similarity between $\mathcal{I}_{\mathcal{S}}$ and other identities in $\mathcal{T}$.

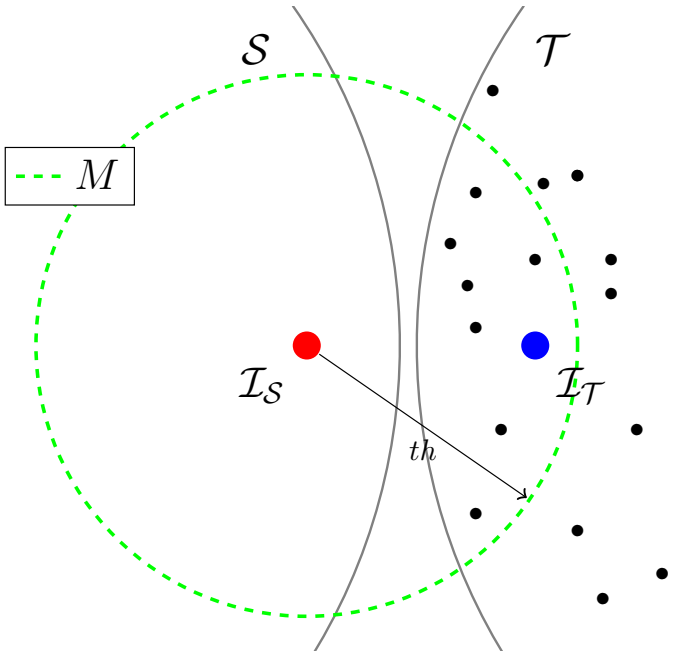


Figure 5.1: Illustration of two domains and the matching set $\mathcal{M}$ of $\mathcal{I}_\mathcal{S}$ in $\mathcal{T}$. The size of the matching set is 8.

we do not know the adversary's threshold choice when estimating the linkability risks of identities.

5.4.5 Anonymity of an Identity

We formalize anonymity in a social media system using the notion of (k, d)-anonymity we introduced in Chapter 4. As we have seen, the notion of (k, d)-anonymity provides a generalization of the classic notion of k-anonymity [103]: (k, d)-anonymity defines the *anonymity set* $\mathcal{A}(d)$ of the target identity $\mathcal{I}_\mathcal{T}$ that contains at least k identities within the target domain $\mathcal{T}$ that have a distance of at most d to $\mathcal{I}_\mathcal{T}$. We briefly recall the corresponding definition.

Definition 23 ((k, d)-Anonymity)**.**
An identity $\mathcal{I}$ is (k, d)-anonymous *in a domain $\mathcal{D}$ if there exists an* anonymity set *$\mathcal{D}' \subseteq \mathcal{D}$ with the properties that $\mathcal{I} \in \mathcal{D}$, that $|\mathcal{D}'| \geq k$ and that all $\mathcal{I}' \in \mathcal{D}$ have* $\mathsf{dist}(\theta_\mathcal{I}, \theta_{\mathcal{I}'}) \leq d$.

We denote with $\mathcal{A}_\mathcal{I}(d)$ the largest anonymity set of $\mathcal{I}$ for a distance of d, and call d its convergence.

Throughout the remainder of the paper we forgo the subscript of the anonymity set $\mathcal{A}(d)$ when we talk about the anonymity set of the target identity $\mathcal{I}_\mathcal{T}$ to keep the notation simple.

5.4.6 Relation of Anonymity and Linkability

In the traditional database setting, anonymity naturally implies unlinkability: notions such as k-anonymity and l-diversity require all identities within an anonymity set to be equivalent. Thus, any source identity cannot be uniquely linked to any target identity in a sufficiently large anonymity set. Ideally, we would want the same to hold in open settings such as social media systems as well: if an identity $\mathcal{I}_\mathcal{T}$ is anonymous in its domain $\mathcal{T}$, it should also be difficult to link it to its matching identity $\mathcal{I}_\mathcal{S}$ since the adversary cannot sufficiently distinguish $\mathcal{I}_\mathcal{T}$ from the other identities in $\mathcal{T}$.

The main question we pose in this paper is whether *the anonymity set size of the target identity $\mathcal{I}_\mathcal{T}$ provides a good assessment of the difficulty of successfully linking the source identity $\mathcal{I}_\mathcal{S}$ to $\mathcal{I}_\mathcal{T}$, i.e., does a large anonymity set imply a large matching set?* Using the notions we introduced in the previous section, our goal is therefore to investigate whether the $\mathcal{I}_\mathcal{T}$'s anonymity, as estimated by its (k, d)-anonymity, can be used to estimate the size of the $\mathcal{I}_\mathcal{S}$'s matching set $\mathcal{M}$.

5.5 Assessing Linkability Risks using Anonymity

We investigate two different scenarios in which we use an identity's anonymity to assess its linkability across social media systems. In the first scenario, we assume that we do not know the adversary's matching strategy (i.e., we do not know the threshold he chooses to link identities – see Section 5.4.3) and we do not know the matching identities of users in other social media systems. Our goal is to see whether the relative anonymity of identities in the social media system can be used to derive a relative linkability measure that informs the users about their linkability risks. In the second scenario we assume the attacker is targeting a particular user and hence, we can combine the (k, d)-anonymity of the $\mathcal{I}_\mathcal{T}$ as well as knowledge about its matching identity $\mathcal{I}_\mathcal{S}$ to develop an *absolute linkability measure*.

5.5.1 Relative Linkability Measure

Context With the relative linkability measure, we want to identify those identities within a domain that are most susceptible to being successfully linked to their matching identities in other domains. Intuitively, and without knowledge about the matching identity, this mostly depends on the uniqueness of an identity within a domain: observe within the same domain that an identity either (a) is very unique and therefore easily identifiable, or (b) blends well into the crowd and therefore has good anonymity.

The notion of (k, d)-anonymity we introduced in Section 5.4.5 essentially captures the uniqueness of the target identity $\mathcal{I}_\mathcal{T}$ in the target domain $\mathcal{T}$. Our hope is that by ranking identities by their anonymity sets, we get a relative assessment of an identity's linkability compared to other identities within the same domain. Against a rational adversary that tries to maximize the number of correct matchings he achieves between two domains, such a relative ranking provides insight into which identity is more likely to be matched first by the adversary.

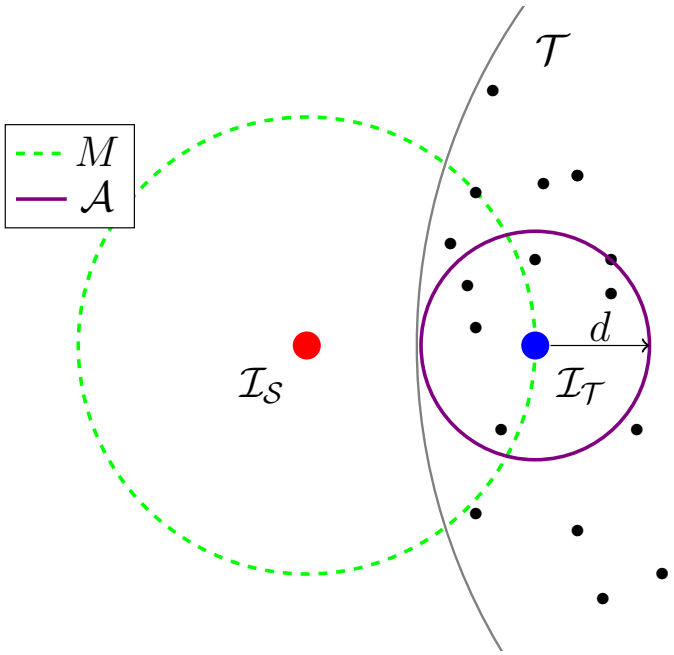


Figure 5.2: Illustration of the anonymity set $\mathcal{A}$ and the matching set $\mathcal{M}$.

Since (k, d)-anonymity has two parameters, we have two options to generate a suitable ranking to predict the relative linkability of user within a domain. The first is to rank identities by their anonymity set size: for a given convergence value d, we compute, for all identities $\mathcal{I} \in \mathcal{T}$, the anonymity set $\mathcal{A}(d)$ and rank the identities by its size. The second option is to rank identities by the convergence of their anonymity sets: here, we fix the anonymity set size k and determine the required convergence value d to achieve k. The identities are then ranked by Independent of how we approach this ranking, the linkability assessment of a specific identity is then derived from its rank: the relative linkability measure thus tells each identity how linkable it is compared to other identities in the same domain.

However, at this point, we do not have any additional information that would support the choice of any specific value for d or k. Instead, we propose a ranking scheme that combines the rankings computed for multiple values of d or k to generate an overall consistent ranking. In the following, we describe this consistent ranking scheme for ranking by anonymity set size. The algorithms can be easily adopted similarly for ranking by convergence.

Consistent Ranking of Identities Given a set of convergence values $\mathbb{D}$ (in our evaluation, we choose all convergence values between 0 and 1, in $\frac{1}{1000}$ steps, since the Jensen-Shannon divergence is bounded by these values), we compute for all identities $\mathcal{I} \in \mathcal{T}$ and for all convergences $d \in \mathbb{D}$ the maximum anonymity set $\mathcal{A}_{\mathcal{I}}(d)$ and rank each identity by the size of these anonymity sets in rank_d. During this ranking, we resolve ties by assigning all identities that have equal set sizes the set of ranks they could occupy. For example, if rank 3 and 4 are not uniquely defined because of a tie

Algorithm 1 Consistent ranking of identities.

Require: Consistent_Ranking($\mathcal{T}$, $\mathbb{D}$)
for $d \in \mathbb{D}$ **do**
  for $\mathcal{I} \in \mathcal{T}$ **do**
    compute $\mathcal{A}_{\mathcal{I}}(d)$
  sort all $\mathcal{A}_{\mathcal{I}}(d)$ into list $\mathcal{L}$
  for $\mathcal{I} \in \mathcal{D}$ **do**
    $\mathsf{rank}_d(\mathcal{I}) = \mathsf{fillingCompRank}(\mathcal{A}_{\mathcal{I}}, \mathcal{L})$
$G = (V = \mathcal{T} \cup \{1, .., |\mathcal{T}|\}, E = \mathcal{T} \times \{1, .., |\mathcal{T}|\}, w)$ with $\forall e \in E : w(e) = 0$
for $d \in \mathbb{D}$ **do**
  for $\mathcal{I} \in \mathcal{T}$ **do**
    $w((\mathcal{I}, \mathsf{rank}_d(\mathcal{I})))+ = 1$
compute maximum weight matching M on G
for $(\mathcal{I}, r) \in M$ **do**
  $\mathsf{rank}^*(\mathcal{I}) = r$
return rank^*

between two identities, both will be assigned the set of ranks $\{3, 4\}$. This procedure that we call fillingCompRank corresponds to a standard competition ranking with filling up the gaps afterwards.

Next, we construct a bipartite graph $G = (V = \mathcal{T} \cup \{1, .., |\mathcal{T}|\}, E = \mathcal{T} \times \{1, .., |\mathcal{T}|\}, w)$ between all identities and their rankings. The weight of an edge $(\mathcal{I}, r)$ in the bipartite graph corresponds to the number of times $\mathcal{I}$ was ranked at rth position in the rank_d rankings.

The final ranking is then determined by the maximum weight matching on the bipartite graph. This ranking scheme takes into account how large the anonymity sets of identities are and also how quickly they grow for varying values of d. A pseudo-code implementation of this algorithm can be found in Algorithm 1.

In the experimental evaluation in Section 5.7.4, we evaluate this consistent ranking method in practice.

5.5.2 Absolute Linkability Measure

Context Contrary to the relative linkability measure, we now make additional assumptions about the adversary: we consider a different scenario in which we know which matching identities $\mathcal{I}_{\mathcal{S}}$ and $\mathcal{I}_{\mathcal{T}}$ the adversary wants to link. For the absolute linkability measure, we include additional information about the source identity $\mathcal{I}_{\mathcal{S}}$ to produce a targeted estimate of linkability. Our goal is to estimate how many identities in the target domain $\mathcal{T}$ match the source identity $\mathcal{I}_{\mathcal{T}}$ at least as well as the matching target identity $\mathcal{I}_{\mathcal{T}}$, i.e., we want to estimate the size of the matching set $\mathcal{M}$.
A first, simple approach to include information about the source identity $\mathcal{I}_{\mathcal{S}}$ in our linkability assessment is to choose the convergence d of the anonymity sets $\mathcal{A}(d)$ as the distance of source and target identity, i.e., $d = \mathsf{dist}(\mathcal{I}_{\mathcal{S}}, \mathcal{I}_{\mathcal{T}})$. Through this, we capture all identities in the neighborhood of $\mathcal{I}_{\mathcal{T}}$ that can potentially appear in the matching set. While other identities, which are not in $\mathcal{A}(d)$, will still appear in the matching set, considering $\mathcal{A}(d)$ might potentially allow us to provide a lower bound estimate on the size of the matching set. However, in some cases, the anonymity set $\mathcal{A}(d)$ will not

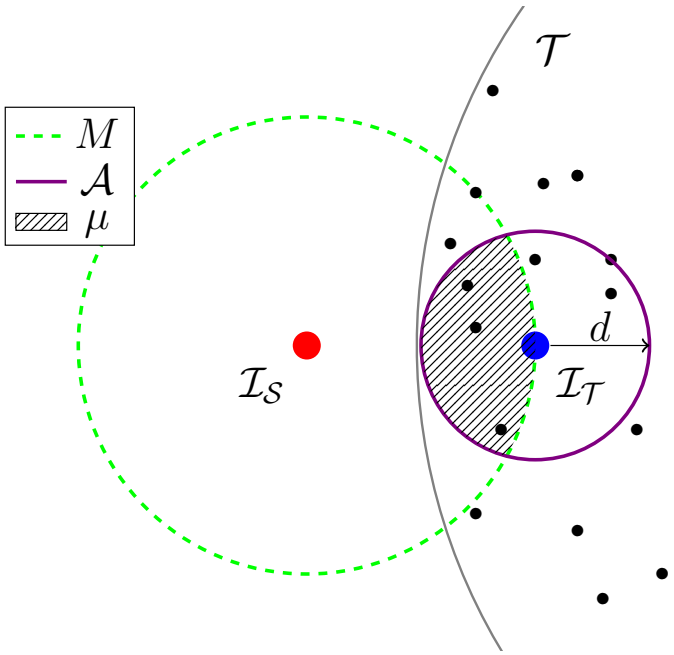


Figure 5.3: Illustration of the anonymity set $\mathcal{A}$, the matching set $\mathcal{M}$, and the local matching set μ.

approximate the size of the matching set $\mathcal{M}$ well: $\mathcal{A}(d)$ might be distributed in such a way that all identities within $\mathcal{A}(d)$ have a distance $d' \geq d$ to the source identity $\mathcal{I}_\mathcal{S}$, and thus $\mathcal{M} \cap \mathcal{A}(d) = \emptyset$. In the illustration in Figure 5.3, this would correspond to the hypothetical case where all identities within $\mathcal{A}$ are outside the matching set $\mathcal{M}$.

Therefore, instead of directly estimating the size of the matching set $\mathcal{M}$ with the anonymity set $\mathcal{A}(d)$, we use the *local matching set* μ, which is the intersection between $\mathcal{M}$ and $\mathcal{A}(d)$ to estimate the size of $\mathcal{M}$.

Definition 24 (Local Matching Set)**.**
Let $d = \mathsf{dist}(\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})$*. Then the* local matching set μ *of the source identity* $\mathcal{I}_\mathcal{S}$ *matching against a target identity* $\mathcal{I}_\mathcal{T}$ *is defined by* $\mu = \mathcal{M} \cap \mathcal{A}(d)$*.*

We illustrate the relation between the matching set $\mathcal{M}$, the anonymity set $\mathcal{A}(d)$ and the local matching set μ in Figure 5.3. Setting the convergence d of the anonymity set to the distance of the matching identities allows us to capture a large part of the identities from the matching set in our local matching set.

5.6 Reddit Data Set

We use Reddit [104] to study the relationship between anonymity and linkability in social media systems. We utilize the same data set already described in Section 4.6.2 Since we aim to asses the risk of linking the identities of the same user across different communities, it is crucial to have ground-truth on matching identities. As already discussed earlier, We opportunistically use Reddit's subreddit structure to obtain such

ground-truth: we treat each subreddit as its own (virtual) domain, and assume that each user has a separate identity in each subreddit. This way, we easily obtain the ground-truth on matching identities, because each user has the same pseudonym across all subreddits. Overall, our data set contains about 2.75 million of such identities.

We apply the same filtering steps already discussed in Section 4.6.2 to avoid noise due to the lack of data: we perform our evaluation only on identities that have at least 100 comments and that belong to a subreddit with at least 100 profiles. Through this, we make sure that (a) each identity provides a sufficient amount of comments to model them (a similar approach has been taken in previous work on author identification as well [77]) and (b) there are sufficient identities within a domain to analyze the distribution of anonymity sets. Furthermore, we dropped the three largest subreddits from our data set to speed up the computation. After this filtering, we retain a data set that contains 15 million comments contributed by $58,091$ different identities that belong to $37,935$ different users in $1,930$ different subreddits. Details about the distribution of identities over the subreddits can be found in supplementary material available online [5].

5.7 Reddit Evaluation

In this section, we evaluate the utility of (k, d)-anonymity to assess te risk of user's identities to be linked across social media systems. We first characterize the size of matching sets and anonymity sets in our Reddit data set. We then investigate whether the relative and absolute linkability measures we proposed in Section 5.5 are a good estimator for the linkability risks, i.e., the matching set size of identities.

Note that, for simplicity, all graphs in this section are based on the source subreddit *news* and the target subreddit *worldnews* if not explicitly mentioned otherwise. During our evaluation process, we also considered other pairs of subreddits for which we provide the same kind of diagrams in supplementary material available online [5]. For each claim, we also provide general graphs summarizing over the whole data set and showing that the results hold across other subreddits as well.

5.7.1 Identity Model Instantiations

As mentioned in Section 5.4.2, we evaluated our measures using various other identity model instantiations. More precisely, we instantiated the identity models not only using (1) unigram frequencies, but also using (2) unigram based indicator vectors, (3) term frequency-inverse document frequencies (TFIDFs), and (4) disjoint author-document topic models [97]. While the first two instantiations do not incorporate the distribution of words within a subreddit, the latter two instantiations were specifically used to separate words belonging to the general topic of a subreddit from author specific language.

For each of these instantiations choices, we evaluated both, the relative and the absolute linkability measures using two different distance/similarity metrics, namely the (a) Jensen-Shannon divergence and (b) Cosine similarity. Our experiments showed that the choice of the distance metric mainly results in a shift of the similarities (and

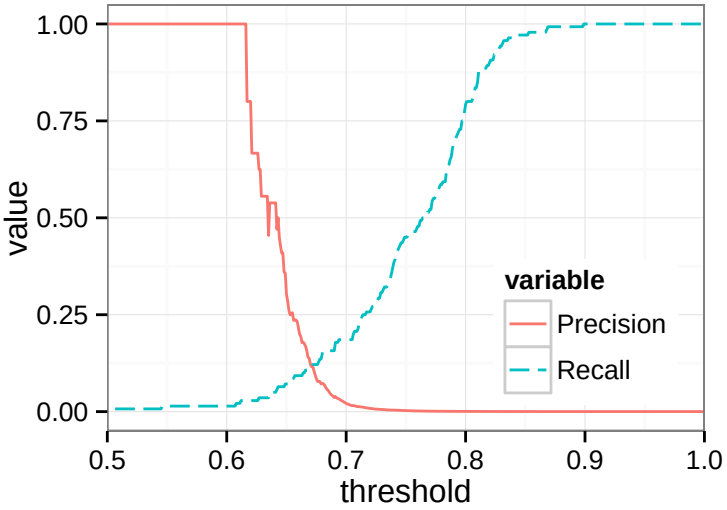


Figure 5.4: Precision and recall tradeoff for matching identities from subreddit *news* to *worldnews*.

hence a shift of the thresholds) without affecting the precision, the recall and the general take-aways. Moreover, while the conclusions drawn from the experiments remained the same for all instantiations of the identity models, the unigram frequency and the TFIDF approaches provided the best, albeit very similar, results with respect to both our estimations and the adversary's linkage attack. Thus, we will, in the following, focus on the results obtained using unigram frequencies for our identity model and the Jensen-Shannon divergence as our similarity metric.

5.7.2 Characterization of Matching Sets

In Section 5.4.3, we explained that a rational adversary tries to correctly match as many identities as possible. To this end, the adversary needs to choose an appropriate threshold on the likelihood that two identities belong to the same user to consider two identities as matching. If the adversary has access to a small ground-truth set (which is the assumption that many previous works in this area make) then he can choose the threshold by analyzing the tradeoff between precision (how many of the identities linked are true matching identities) and recall (how many identities are linked out of all true matching identities). In this paper, we assume that the adversary takes the distance between identities as the likelihood measure. Figure 5.4 depicts both the precision and recall of an adversary for varying thresholds for matching identities in the *news* subreddit to identities in the *worldnews* subreddit.

Since the choice of threshold will of course impact the size of the matching sets we plot, in Figure 5.5, the median and mean size of the matching sets depending on the threshold. For example, the median matching set size for a threshold of 0.8 is 37.

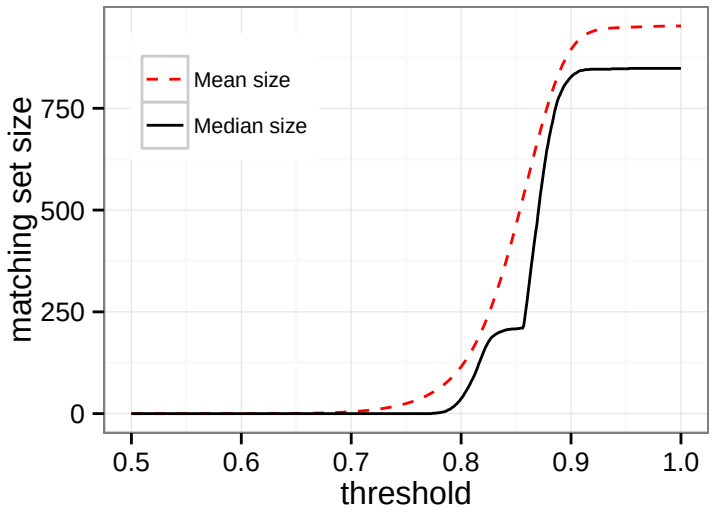


Figure 5.5: Median and mean matching set sizes of the adversary depending on the chosen threshold (for matching identities from subreddit *news* to *worldnews*).

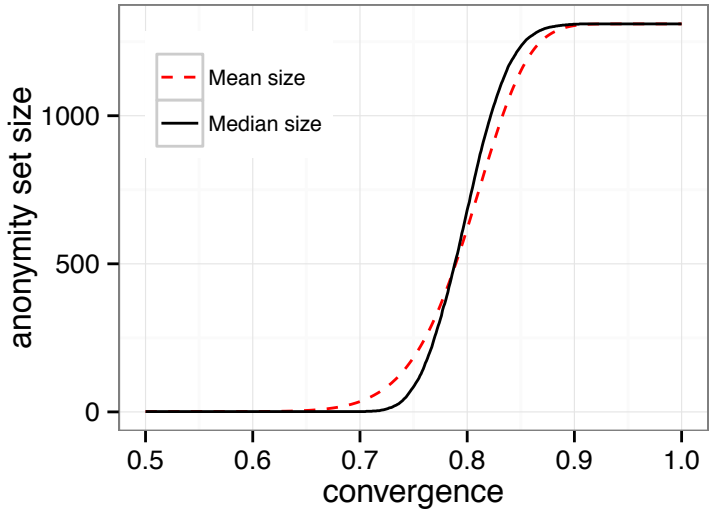


Figure 5.6: Median and mean anonymity set sizes when varying the convergence d (subreddit *worldnews*).

5.7.3 Characterization of Anonymity Sets

Since the notion of anonymity sets lays the foundation of our two linkability measures, we first have a closer look at its characteristics in our data set. Figure 5.6, plots the median and mean size of the anonymity set (for the subreddit *worldnews*) depending on the convergence d. For example, the median anonymity set size for a convergence of 0.8 is 37, which very is similar to the median matching set size for the same threshold.

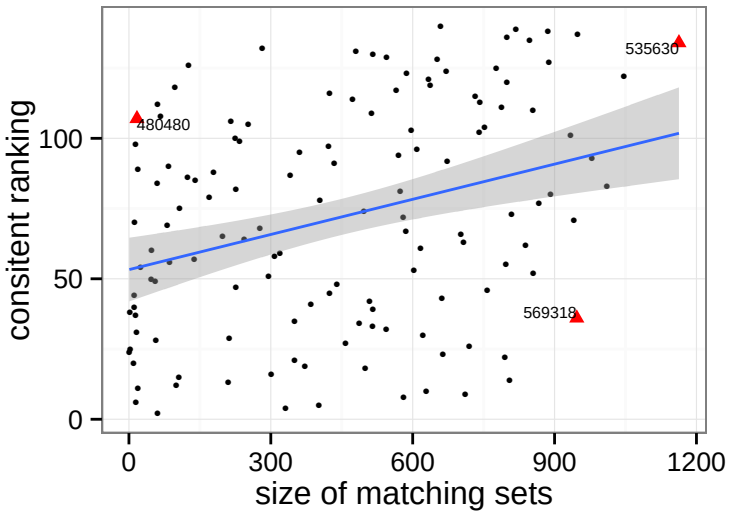


Figure 5.7: The consistent ranking (over all convergences) compared to the size of the corresponding matching sets for an adversary's threshold of 0.8 (subreddit *news* to *worldnews*).

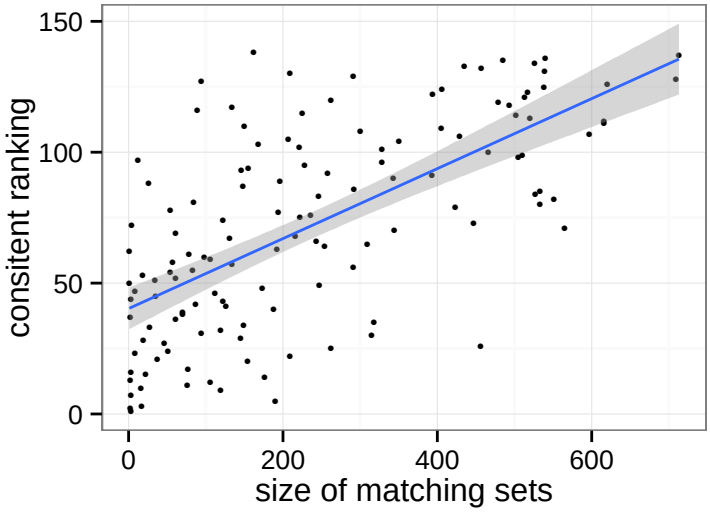


Figure 5.8: The consistent ranking (over all convergences) compared to the size of the corresponding matching sets for an adversary's threshold of 0.85 (subreddit *pics* to *wtf*).

5.7.4 Assessing the Relative Linkability Measure

Remember that, in Section 5.5.1, we introduced the relative linkability measure to identify, within a domain, the identities that are most at risk of being linked to their matching identities in other domains. In this section, we investigate whether the relative

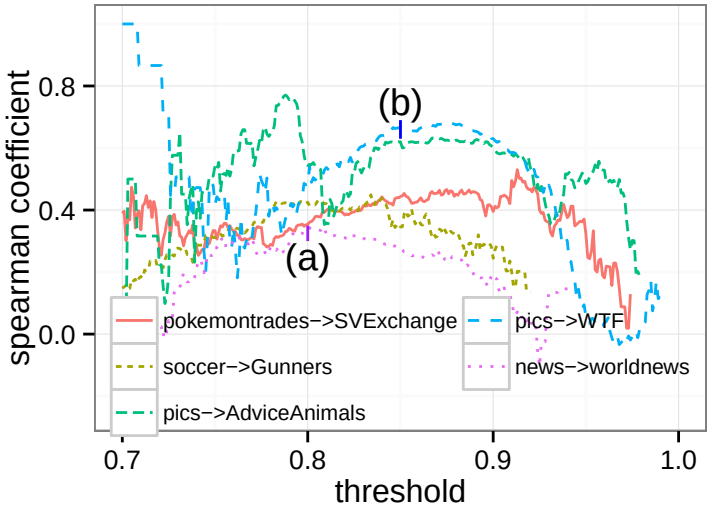


Figure 5.9: Spearman's correlation coefficient between the consistent ranking and the size of the matching sets for different adversary thresholds and different subreddits.

linkability measure is a good estimate of linkability risks. Thus, our goal is to investigate to which degree the consistent ranking provided by the relative linkability measure correlates with the matching set size $\mathcal{M}(th)$.

Note that since this measures relies only on a minimal amount of information, i.e., it only takes into account the similarities between identities in a single domain and does not take into account the matching identities of a user, we do expect the approximation not to hold in all cases.

Figures 5.7 and 5.8 depict the correlation between the size of an identity's matching set (for a particular threshold) and the rank of the identity for two different pairs of subreddits. In both figures, we see a positive correlation between the consistent ranking and the size of the matching set, however, Figure 5.8 presents a better correlation than Figure 5.7.

To illustrate how the correlation depends on the threshold chosen by the adversary and the pairs of subreddits considered, Figure 5.9 depicts the Spearman correlation coefficient between the consistent ranking and the size of the matching set for various thresholds and multiple subreddit pairs. For reference, the thresholds for the previous figures are also annotated. The figure shows that there is a positive correlation between the consistent ranking and the size of the matching set for other pairs of subreddits as well. However, for all the thresholds considered, the correlation is not very strong in general.

Furthermore, in Figure 5.7 we can see that there are many points that are far from the regression model. There are identities with a high rank that have a small matching set, and there are identities with a low rank that have comparatively large matching sets. While the consistent ranking overestimates the linkability risk of the identity in the bottom right corner which might not be so problematic; it underestimates the linkability risk for the identity in the top left corner which is really problematic because

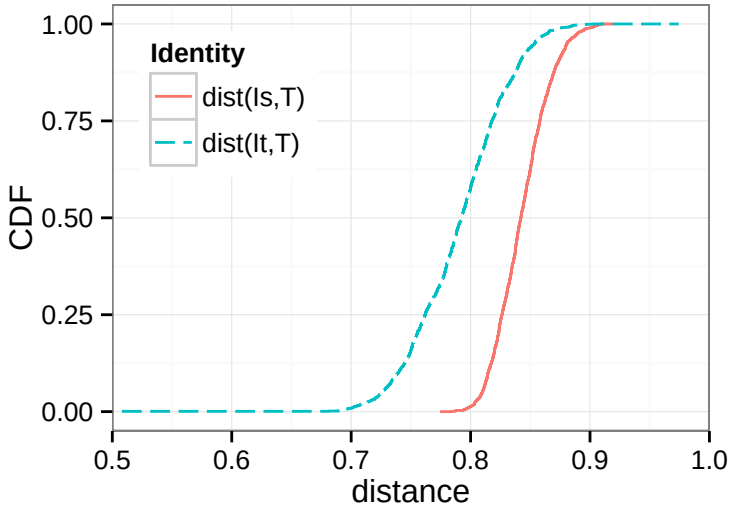


Figure 5.10: CDFs of distances from $\mathcal{I}_\mathcal{T}$ and $\mathcal{I}_\mathcal{S}$ to the target subreddit $\mathcal{T}$. Case of underestimation of the linkability risk (for user id 480480).

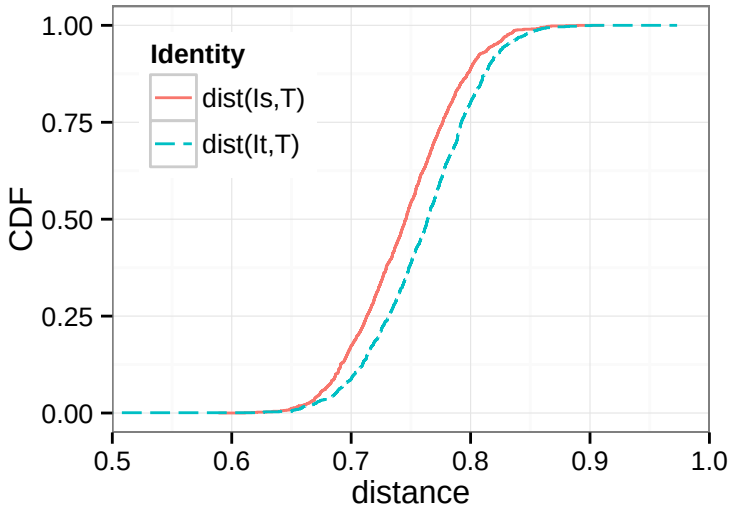


Figure 5.11: CDFs of distances from $\mathcal{I}_\mathcal{T}$ and $\mathcal{I}_\mathcal{S}$ to the target subreddit $\mathcal{T}$. Case of good estimation of the linkability risk (for user id 535630).

it makes the identity subject to a false sense of anonymity.

To investigate why in some cases the consistent ranking estimates well the size of the matching set while in other cases it overestimate or underestimates it, we further investigate the three highlighted identities. To this end, we analyze the relation between the distances $\mathsf{dist}(\mathcal{I}_T, \mathcal{T})$ from the target identity $\mathcal{I}_T$ to the target subreddit $\mathcal{T}$ and the distances $\mathsf{dist}(\mathcal{I}_S, \mathcal{T})$ from the source identity $\mathcal{I}_S$ to our target subreddit $\mathcal{T}$. We present the CDFs of these distances for the three identities that have been highlighted

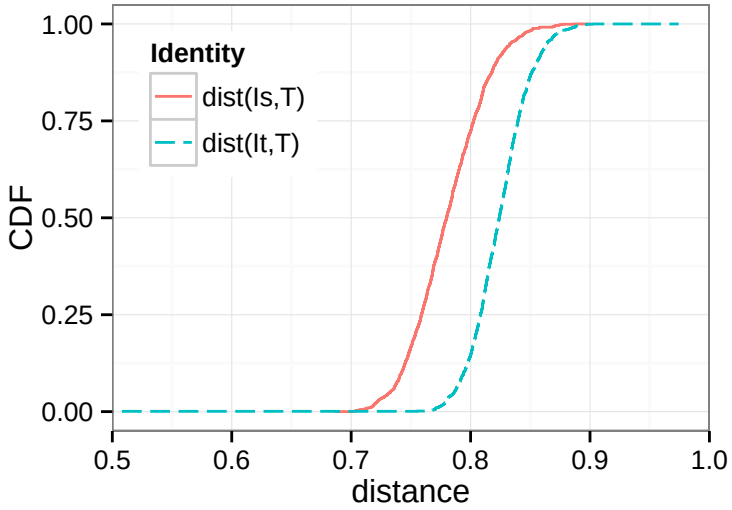


Figure 5.12: CDFs of distances from $\mathcal{I}_\mathcal{T}$ and $\mathcal{I}_\mathcal{S}$ to the target subreddit $\mathcal{T}$. Case of overestimation of the linkability risk (for user id 569318).

in the previous figure (a case of underestimation, a case of good estimation and a case of overestimation). In both Figure 5.10 (underestimation) and Figure 5.12 (overestimation), the distributions are rather dissimilar while in Figure 5.11 (good estimation) the distributions are rather similar. In the first case, the distances to identities in the same subreddit (from $\mathcal{I}_\mathcal{T}$ to $\mathcal{T}$) are smaller than those when matching from the outside (from $\mathcal{I}_\mathcal{T}$ to $\mathcal{T}$) which leads to large anonymity sets and small matching sets, which leads in turn to the false sense of anonymity for that particular identity. In the second case, the distances to identities in the same subreddit are larger, which consequently leads to an overestimation of the identity's linkability risk. Thus, the accuracy of the (k, d)-anonymity to estimate the linkability risk depends on how an identity $\mathcal{I}_\mathcal{T}$ is placed with respect to other identities in the domain (as measured by the similarity between them) as well as on how far the matching identity $\mathcal{I}_\mathcal{S}$ is from identities in the target domain. The absolute linkability measure takes exactly this into account.

5.7.5 Assessing the Absolute Linkability Measure

The absolute linkability measure, as explained in Section 5.5.2, aims to assess the linkability risk if an adversary targets a particular user. Thus, the goal of this section is to investigate whether the anonymity set $\mathcal{A}(d)$ where $d = \mathsf{dist}(\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})$ and the corresponding local matching set μ estimate reliably the size of the matching set $\mathcal{M}(th)$ for a threshold $th = d$.

Anonymity Set Figures 5.13 depict the correlation between the size of the anonymity set $\mathcal{A}(d)$ and the size of the matching set $\mathcal{M}(d)$ for matching identities between subreddits *news* and *world news*. Note that, compared with the previous section where we had the same th for all pairs of identities, here, for each pair or identities we compute

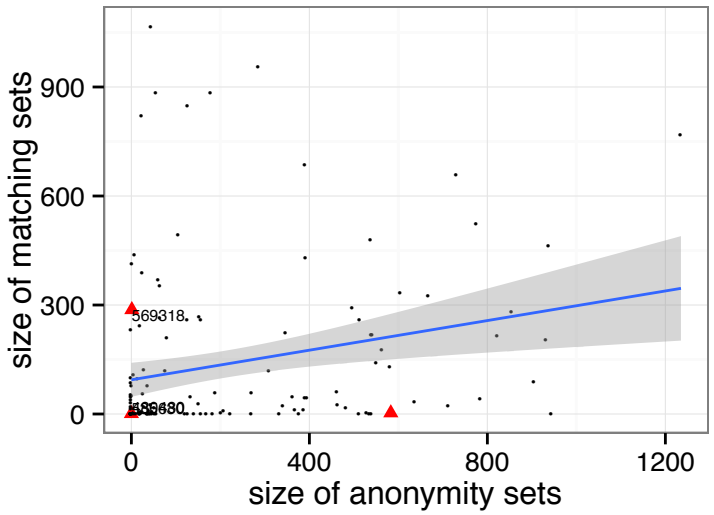


Figure 5.13: Size of the anonymity set $\mathcal{A}(d)$ compared to the size of the matching set $\mathcal{M}(d)$ where $d = \mathsf{dist}(\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})$ for each pair of identities (subreddit *news* to *worldnews*).

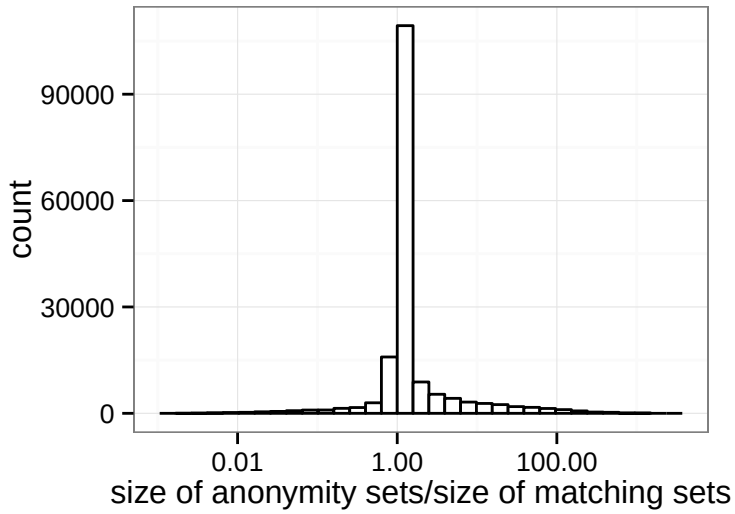


Figure 5.14: The fraction of the anonymous set sizes to the corresponding matching set sizes $\frac{|\mathcal{A}(d)|}{|\mathcal{M}(d)|}$ where $d = \mathsf{dist}(\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})$ over all pairs of subreddits in our data set.

$\mathcal{A}(d)$ and $\mathcal{M}(d)$ where $d = \mathsf{dist}(\mathcal{I}_\mathcal{S}, \mathcal{I}_\mathcal{T})$. The Spearman's correlation coefficient for this plot is 0.41, comparable to the values we obtained in the previous section.

To check the general correlation of anonymity sets and matching sets in other subreddits, Figure 5.14 inspects the ratio of anonymity set sizes and matching set sizes $\frac{|\mathcal{A}(d)|}{|\mathcal{M}(d)|}$ on our whole data set. When over-approximating the risk of an identity the anonymity set size is small compared to the size of the matching set, resulting in a

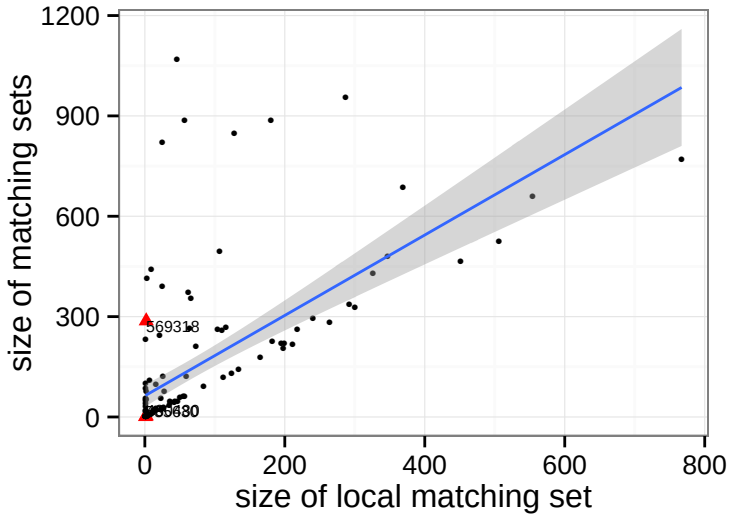


Figure 5.15: Comparison of the sizes of local matching sets and matching sets.

ratio < 1. Conversely, when under-approximating the risk, the ratio is > 1. Note that the x-axis is plotted in log scale to allow us to display the tail ends of the distribution. We can see a clear peak in the area around 1: for at least 71% of all cases, the fraction $\frac{|\mathcal{A}(d)|}{|\mathcal{M}(d)|}$ lies in the interval $[0.8, 1.2)$. Thus, for most pairs of subreddits there is correlation between $\mathcal{A}(d)$ and $\mathcal{M}(d)$. However, the anonymity set size suffers from the same drawback as the relative linkability measures, it underestimates the linkability risk for 41.2% of identities in our data set which makes the anonymity set size not a reliable measure of linkability risks.

Local Matching Set Figure 5.15 depicts the size of an identity's local matching set compared to the size of the adversary's matching set for our exemplary pair of subreddits. Clearly, except for a few outliers, both sizes positively correlate. The few outliers only provide an over-approximation of the identity's risk: While the local matching set is small and the identity does not seem to blend into the crowd, the matching set is large and thus the identity cannot easily be linked.

The more intriguing question, however, is how accurately the local matching set estimates the matching set. To this end, we analyze the ratio $\frac{|\mu|}{|\mathcal{M}|}$ between both sizes on our whole data set in Figure 5.16. If both set sizes coincide, the ratio yields 1, whereas inaccurate estimations of the matching set size will result in a ratio towards 0. We can see that for the vast majority of identities, both sets coincide or at least are very similar: In at least 74% of the cases, the local matching set has at least $0.8 \cdot |\mathcal{M}(d)|$ elements.

The local matching set is a much better linkability risk measure than the anonymity set because it takes into account the structure of the identity space, i.e., the positioning of identities with respect to each other based on the similarities between them. Figure 5.17 plots the correlation between the anonymity set size and the local matching

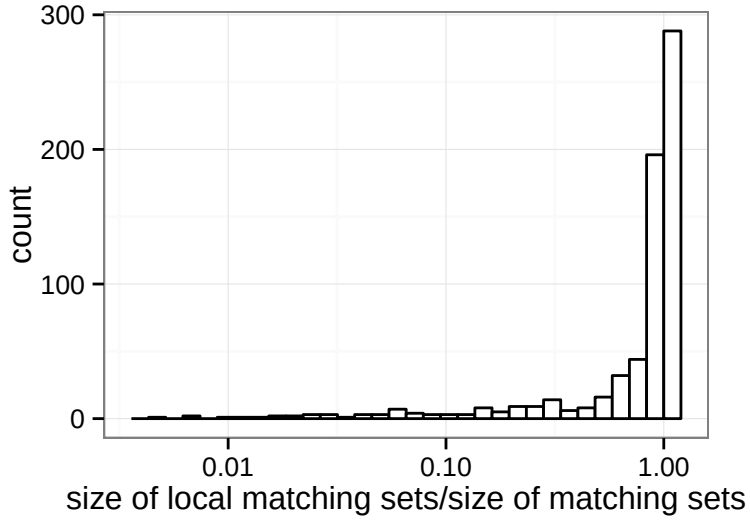


Figure 5.16: The conformity of local matching set sizes to the corresponding matching set sizes $\frac{|\mu|}{|\mathcal{M}(d)|}$ over all pairs of subreddits in our data set.

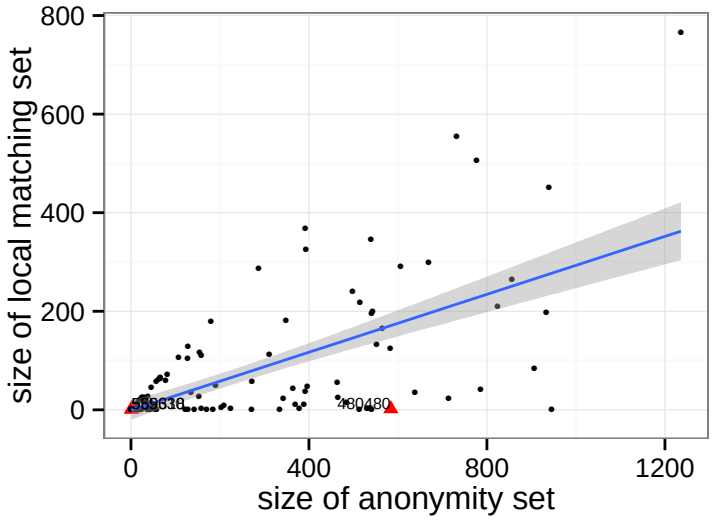


Figure 5.17: Size of the local matching set compared to the size of the anonymous subset.

set size. They do not correlate perfectly because the identities inside a anonymity set are not distributed uniformly. This makes the anonymity set alone a bad estimate for linkability, because only a fraction of the identities in the anonymity set may be relevant for the matching of two identities. The local matching set, on the other hand, only contains identities that also appear in the matching set $\mathcal{M}$ and thus provide a lower bound for its size.

5.7.6 Discussion

The insights obtained through the experimental evaluation are twofold: First, the consistent ranking by anonymity set size we derived in Section 5.5.1, while showing some positive correlation with the matching set size, does not provide a sufficient assessment of the linkability of a user's identities across social media platforms. Second, extending (k, d)-anonymity with information about the matching identity results in local matching sets that provide a good approximation of the absolute risk for matching identities to be linked.

Insufficiency of Anonymity By our first insight, we can conclude that only considering the anonymity of an identity within domain is not sufficient to assess the likelihood of linking matching identities across domains. This is contrary to the traditional database setting, where we have a strong relation between anonymity in linkability due to the pre-defined and restricted number features (i.e. columns in the database) and the required exact equality for anonymity [86].

Such a results was to be expected: the linking process itself utilizes much more information than what is used for determining the anonymity of an identity within a community. As discussed by Goga et al. [39], properties that allow for a successful matching (for instance availability and consistency of identity attributes) depend on both source and target identity. Therefore focusing only on the target identity and its anonymity would not be sufficient to provide a good assessment of linkability.

Absolute Linkability Measure As a solution to this problem, we propose using the size of the local matching set to assess linkability risks: this approach takes into account the source identity to determine the number of identities within the target identity's community that actually allow her to hide herself from the source identity. Our evaluations show that the absolute linkability measure based on local matching sets provides a much better estimation of linkability risks than the simple use of anonymity set sizes.

In practice, using local matching sets presents an approach to assessing linkability risks soundly even if we only consider a subset of the whole domain: by their definition, anonymity sets, local matching sets and matching sets can only grow by increasing the number of identities within a domain. Since, by our evaluations, the local matching set size provides a good approximation of the matching set size, even in very large social media systems with millions of users, it is sufficient to only determine the local matching set size on a subset of the whole domain to provide a meaningful linkability risk assessment. Further increasing the number of considered identities can only increase local matching set and matching set size due to their monotone nature. We therefore only need to gather as much data as is necessary to achieve the linkability assessments that we are satisfied with.

Defensive Mechanisms From our evaluations, we can also infer directions for potential defensive mechanisms against linkage. In general, users should try to increase the distance between their matching identities since this also increases the matching set size, and therefore decreases the potential linkability.

Furthermore, users should try and avoid exhibiting unique features. We observed in our evaluations for the absolute measure in Section 5.7.5 that the anonymity set $\mathcal{A}$ of an identity alone is not a good assessment of linkability due to the potentially uneven distribution of identities in $\mathcal{A}$. Such an uneven distribution can be caused by unique attributes that are exhibited by the source and target identity, but not by other identities in the target domain. In Chapter 4 we captured this under the notion of critical attributes. In related work, Goga et al. [39] capture this under the notion discriminability of attributes.

Limitations In our evaluations, we represent identities with a unigram identity model based on the comments users post on Reddit. As discussed in Section 5.4, users also share other types of content, such as audio, video and text content that can be analyzed in a much more elaborate manner. Including more features of user-content into our analysis will induce different identity models with a (possibly different) corresponding distance measure. We expect our anonymity measures to be applicable to such different settings, since they rely on the relation between anonymity set and (local) matching sets that also hold for other metric spaces than the one considered in this paper.

5.8 Conclusion & Future Directions

In this section, we investigated whether anonymity within a social media system is sufficient to protect against the linking of a user's identities across social media sites. To this end, we presented two novel approaches to estimating the linkability of identities by their anonymity within their communities. The relative measure provides a ranking of identities only based on their intra-domain distances to other identities in the same domain. The absolute linkability measure, on the other hand, seeks to directly estimates the size of the matching set (i.e. the number of within a domain that the adversary considers potentially matching a given source identity). To this end we consider, we consider the size of the anonymous subsets alone to estimate the matching set size, as well as introduce local matching sets to use information about the matching identities of the same user in a different social media domain to provide a better estimate of linkability.

We empirically evaluate both measures on a data set of user-generated text content collected from Reddit. We show that, on the one hand, the relative measure that relies on anonymity alone is not sufficient for assessing linkability. The absolute measure, on the one hand, also does not provide a meaningful estimation of linkability if we only rely on anonymity sets. On the other hand, it does provides a meaningful assessment of linkability if we use local matching sets to estimate the size of the matching set. In particular, the absolute linkability measure is also suitable for application in practice: it does not rely on information about all identities within a social media system, but instead can be evaluated on a local subset of all identities, thus greatly improving the computational tractability of assessing linkability in very large social media systems.

These results show that, in contrast to traditional privacy settings such as statistical databases where anonymity alone is sufficient to also provide unlinkability, in open

settings it is important to take into account which identities are being matched to produce a meaningful linkability risk estimate.

In addition to the directions discussed in Section 5.7.6, we consider the following direction important for future work: in practice, social media systems have an ever-changing set of identities that participate, while in this work we consider a static set of identities. Therefore, an efficient method for computing and updating anonymity sets needs to be developed to deal with the dynamically changing nature of social media systems.

6

Reconciling Privacy and Utility in Continuous-time Diffusion Networks

6.1 Motivation

Social networks like Facebook or Twitter are used daily by billions of users to communicate and interact with their peers. In particular, their networked structure allows users to easily and quickly share any kind of information with a large audience. This, however, also has its drawbacks: once shared, the transitive diffusion of information in such networks can cause information to quickly spread through the whole network (might even cause it to become viral), and thus reach users in the network that the information was not originally intended to reach.

Clearly, the simplest solution to avoiding such issues is to not share any information at all. This, however, violates the utility expectation of the user in the social network: sharing (potentially sensitive) information with chosen peers is often a basic necessity to enable interaction with said peers. The question therefore naturally arises how one could approach controlling the propagation of information in social networks while at the same time fulfilling utility expectation by the users.

6.2 Problem Description

Diffusion models, such as the independent cascade model proposed by Goldenberg et al. [40] or the continuous-time diffusion model proposed by Gomez-Rodriguez et al. [43], model the global information diffusion in networked system. They have successfully been used for the analysis and prediction of diffusion processes in various domains, for instance viral marketing [17, 27, 110], epidemiology [107], information propagation [3, 44], and influence estimation and maximization in social networks [45, 57].

The issues of controlling the propagation of information in such diffusion processes while maintaining utility is, as of yet, still unexplored. A comprehensive approach to this issue is, however, necessary for enabling users to enforce their privacy in an increasingly digitalized and connected world.

6.3 Contributions

In this paper, we develop a novel approach to controlling the propagation of information in networked social media systems while at the same time satisfying user-specified utility requirements. Our results provide a formal foundation for simultaneously approaching the issue of both privacy and utility in information propagation. In particular, we provide insights into the algorithmic complexity of optimally controlling information propagation under privacy and utility constraints.

Privacy Policies for Diffusion Networks

We leverage the continuous-time diffusion model [43] as a representation of the information diffusion process in real social networks. Based on these representation, we define two types of *propagation policies* that reconcile privacy and utility requirements: *utility-restricted privacy policies* put a lower bound on the number of friends the user wants to share a specific piece of information with while minimizing the *exposure*, i.e.

the expected number of *malicious users* in the network that the shared information reaches. *Privacy-restricted utility policies*, on the other hand, put an upper bound on the allowed exposure while maximizing the number of friends the information is shared with. The continuous time parameter used in the continuous-time diffusion model furthermore allows us to capture the temporal component of controlling information propagation: by including a time threshold in the propagation policies, we allow the user to specify how long a policy should stay valid, therefore further increasing the potential utility.

Algorithmic Tractability of Privacy in Diffusion Networks

After introducing the formal framework in which we define these policies, we investigate the tractability of optimally satisfying both policy types. To this end, we first show that MAXIMUM-k-PRIVACY–the minimization problem that corresponds to utility-restricted privacy policies–is NP-hard even if provided with an oracle that computes the exposure for a given set of friends (a problem which itself is #P-complete [17]). We show the NP-hardness through a reduction from the MINIMUM k-UNION problem that has been shown to be NP-hard by Vinterbo [106]. Since this precludes an efficient and exact algorithmic solution to our problem, we, in the next step, turn towards solving it approximately: we show that our objective function, the exposure, is submodular, and we are therefore confronted with constrained submodular minimization problem. While these have been shown to be very hard to approximate in general [102], we leverage a recently proposed approximation algorithm by Iyers et al. [53] and achieve a constant factor approximation by utilizes the non-zero *curvature* of our submodular objective function.

Similarly, we show that MAXIMUM-τ-UTILITY–the maximization problem corresponding to the privacy-restricted utility policies-is NP-hard as well. We then leverage the constant factor approximation for MAXIMUM-k-PRIVACY to design an approximation algorithm for MAXIMUM-τ-UTILITY and show that this algorithm provides a constant factor approximation as well.

Outline We begin by introducing notions and definitions used throughout the chapter in Section 6.4. We then introduce the formal framework in which we define our propagation policies in Section 6.5.1. In Section 6.6, we then derive the optimization problems that correspond to our propagation policies, before we take a look at the algorithmic complexity and approximation of these optimization problems in Sections 6.7 and 6.8. We discuss limitations and potential extension of our approach together with potential directions for future work in Section 6.9. Finally, we conclude in Section 6.10.

6.4 Background

We briefly introduce basic notions and definitions that we use throughout this and the next chapter.

6.4.1 Information Diffusion Networks

Information diffusion networks are used to model information, influence or disease spreading behavior in large networks. A significant number of works have explored various different approaches to suitably model these issues [40, 45, 57]. The basic definition of a diffusion network, however, remains mostly the same. It boils down to a network of nodes with directed edges that indicate the potential transmission of information, diseases, etc.

Definition 25. *A* diffusion network *is a tuple $N = (V, \alpha)$, where $V = \{v_i\}$ is the set of nodes in the network, and $\alpha = (\alpha_{i,j})$ with $\alpha_{i,j} \geq 0$ is the* transmission matrix *of the network.*

In our case, the set of nodes V represents the users a social network that exchange information. The transmission matrix α provides us with *pairwise transmission rates* from which we derive the likelihood of information transmission from the node v_i to v_j. Throughout the paper, we will denote this with v_i *infecting* v_j. Together, V and α define a directed graph where each $\alpha_{i,j} > 0$ represents an edge between two nodes v_i and v_j along which information can potentially flow.

Diffusion models proposed in the literature mostly differ in how they model the transmission of an infection given the transmission matrix α. For the work presented in this paper, we leverage the continuous-time diffusion model originally proposed by Gomez-Rodriguez [43]. This model generalizes the independent cascade model, originally proposed by Goldenberg et al. [40], with a dependence of the transmission likelihoods on a continuous time variable. This time variable will later allow us to include a temporal component into our propagation policies that enable the user to also consider the temporal criticality of information: if the propagation of information into a set of malicious nodes only has to be controlled until a time t, we can initially share the information with potentially more friendly nodes.

Let a node v_i be infected at time t_i. Then, in the continuous-time diffusion model, the likelihood that node v_j without being infected by node v_i until time t is given by *survival function* $S_j(t \mid t_i, \alpha_{i,j})$. The exact form of this survival function depends on the model type. For instance, in an exponential model, the survival function is given by

$$S_j(t \mid t_i, \alpha_{i,j}) = e^{-\alpha_{i,j}(t-t_i)}.$$

Now, let $I \subseteq V$ be the set of already infected nodes in the network. Given the survival function, independent of its exact form, we can compute the likelihood that v_j survives without being directly infected by any of the nodes in I until time t with

$$1 - \prod_{v_i \in I} S_j(t \mid t_i, \alpha_{i,j}).$$

While the set of nodes V is directly given by the set of actors present in our network of choice, the transmission parameters are usually not pre-defined but need to be inferred from past information diffusion behavior. Gomez-Rodriguez et al. [43] show that these parameters can successfully be learned for a continuous-time diffusion model from real networks and can then be applied to predict future diffusion behavior in the

network. In this paper, we are not interested in the inference of the transmission likelihoods, but instead assume that the inference has already happened and we have an accurate estimate of $\alpha_{i,j}$ for the whole network.

6.4.2 Submodular Set Functions

In this work, we will reduce the question of reconciling privacy and utility in the above defined information diffusion networks to optimization problems that maximize privacy subject to a utility constraint (or vice versa). While our objective functions will turn out to be NP-hard to optimize, we will also see that they are *submodular*.

Definition 26 (Submodular Set Function)**.** *Let Ω be a finite set. A set function $f : 2^\Omega \to \mathbb{R}$ is* submodular *if for all $X, Y \subseteq \Omega$ with $X \subseteq Y$ and for all $x \in \Omega \setminus Y$ it holds that*

$$f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

While there exist several other, equivalent definitions of submodular set functions, the above *diminishing returns* property of submodular set functions will be most helpful for the results developed in this paper.

A simple example for a submodular set function is the total coverage area $\mathcal{A}$ of a set of surveillance cameras in an area: while each camera C_i will have its own coverage area A_i based on range and obstructions in the area, the total coverage area is not simply given as the sum of all A_i. Instead, due to potential overlaps in the A_i's, each camera might only contribute parts of its own coverage area to the total coverage, i.e. has diminishing returns. Finding the optimal positions for a fixed number of k surveillance cameras that maximize the total coverage area is thus a submodular optimization problem. Other examples include joint entropy (or mutual information) of a set of elements, or various graph cuts, including the number of cut edges.

In general, submodularity has become a valuable tool in many maximization problems. While subdomular maximization has been shown to be NP-hard [36] in general, it can be approximated to a constant factor efficiently using the greedy approach [82, 37].

The optimization problems that we encounter throughout the paper will turn out to be constrained, submodular minimization problems. While these can only be approximated efficiently to a polynomial factor in general [102], Iyers et al. [53] recently showed that well-formed submodular objective functions f can be minimized under constraints up to a constant approximation factor that depends on the structure of f. We utilize this result in Section 6.7.3 to provide an efficient approximation of our optimization problems.

6.4.3 Influence Estimation

In order to control information diffusion, we want to bound the number of potentially malicious nodes in the network that eventually are infected by information we shared with a subset of our friends. In the literature, a closely related problem is known as influence estimation. Here, we want to estimate the number of nodes $v_i \in V$ that are

infected within a time window t given that a subset $A \subseteq V$ was initially infected at time 0.

Definition 27 (Influence)**.** *Let $N = (V, \alpha)$ be an information diffusion network. The* influence $\sigma_N(A; t)$ *of $A \subseteq V$ in N within time t is given by*

$$\sigma_N(A; t) = \sum_{v_i \in V} \Pr[t_i \leq t \mid A].$$

Here, $\Pr[t_i \leq t \mid A]$ *is the likelihood that the infection time t_i of node v_i is smaller than t, given that A is infected at time* 0*.*

Exactly computing the influence caused by an initial infection A has been shown to be $\#P$ complete for the independent cascade model by Chen et al. [17], as well as for the continuous-time diffusion model by Gomez-Rodriguez et al. [46]. While the exact influence can therefore not be computed efficiently, Gomez-Rodriguez et al. provide an efficient, randomized algorithm based on graphical models that approximates σ_N to a constant factor.

The typical application of influence estimation in information diffusion networks is to find the most influential initial subset to share information with. Formally, we want to find $A \in V$ that maximizes $\sigma(A; t)$. Gomez-Rodriguez and Schölkopf show that influence maximization in the continuous-time diffusion model is NP-hard [46]. However, they also show that $\sigma(A; t)$ is submodular in A, and thus allows for an approximation up to a constant factor of $(1 - \frac{1}{e})$ using a greedy heuristic.

For our use case, we are interested in exactly the opposite: we want to identify a subset of friendly nodes (with a cardinality lower bound) that minimizes the influence within a set of malicious nodes. In Section 6.7, we leverage the submodularity of $\sigma(A; t)$ to show that our objective function is submodular as well and further investigate the tractability of this minimization problem.

Note that the literature on influence estimation generally assumes access to an influence estimation oracle to focus on the tractability of the actual combinatorial problem. We will follow a similar approach in our tractability analysis presented in this paper.

6.5 Privacy in Diffusion Networks

In the following, we motivate our general approach to privacy in diffusion networks and discuss how we reconcile it with utility. In the literature, the issue of privacy is typically considered in terms of an adversary inferring sensitive information from the data published by the user. For instance, Differential Privacy [30] protects entries in statistical databases from revealing sensitive information by perturbing the sensitive columns in the database, while at the same time ensuring limited utility by carefully calibrating the applied noise [56]. Providing general and provable non-inference guarantees in the style of Differential Privacy is, however, not very practical in the scenario where users openly communicate in social networks and other social media systems: the sensitivity of information often depends on the context in which it is exhibited and the perturbation by adding noise is not really compatible with utility requirements of the

user–we simply cannot protect a user against revealing sensitive, personal information to a global adversary if their goals is to share this information with certain other users for the sake of discussion, finding friends or other goals.

In this paper, we instead consider privacy in terms of controlling the spreading of information within a network of users that share information with each other. Typical examples of such communication networks include, for instance, Twitter, where users can spread information in tweets very quickly by re-tweeting, or Facebook, where, depending on the user's privacy settings, shared information can traverse the network through likes. Our goal is to enable users to share information in social networks in such a way that, ideally, only the intended recipients receive the information.

In the following we formalize privacy and utility requirements for sharing information in social networks by leveraging the diffusion networks (cf. Section 6.4) as a representation of the global information diffusion in such networks.

6.5.1 Privacy Model

We formalize privacy requirements for information propagation through a set of nodes in the diffusion network that should, ideally, not receive a piece of information within a time frame t after the information was initially shared. Given a diffusion network $N = (V, \alpha)$, this set $M \subseteq V$ of *malicious nodes* in the network is chosen by the user for each of their actions. That is, each time the user shares information, they can specify a different malicious node set that should not receive this information.

Example 1. *Alice uses a social network to communicate with her family and friends, but also with her colleagues from work. She now performs an action that contains information she only wants to share with her family. Users in the friends and work group that do not belong to her family should not receive this information. For this specific action, all users in the friends and work group that are not part of the family group would be considered as malicious nodes.*

To control the propagation of information into the malicious node set M, Alice has to carefully choose the initial infection of the diffusion network, i.e. the set of nodes with which she initially shares information with. More formally, in the continuous-time diffusion model, we want to minimize the expected number of nodes in M that receive information shared by a user within a time frame t. This is exactly the influence within the malicious node set M caused by the set of nodes that the user initially shared the information with. In the formalizations we present below, we will call this the *exposure* of the information.

Note that the above formulation does not consider any type of global adversaries that observe information from outside the network structure. In practice, this means that our approach cannot protect against provider or state-level adversaries that potentially directly observe any information shared by the user. Still, our model allows us to soundly estimate the propagation of information in existing social networks where users only share information with other users they are directly connected to (assuming the underlying diffusion network accurately models the diffusion process). While this does not represent all social media platforms, it does encompass some of the biggest

representatives such as Facebook or LinkedIn (assuming the user utilizes the privacy options offered by these platforms).

6.5.2 Utility Model

Clearly, the easiest way to minimize the influence within M is to not share the information with anyone to begin with: if there is not a single node in the diffusion network that is infected, the information can also not further spread throughout the network, and eventually reach a malicious node. However, this also removes any *utility* that the user might want to achieve through their actions. To formalize utility, we also consider the set $F \subseteq V$ of *friendly* nodes with $F \cap M = \emptyset$. This friendly node set corresponds to the nodes in the network which the user wants to share information with. As in the malicious node case, this friendly node set may also be re-defined for every action performed by the user. In Example 1, the friendly node set would correspond to all nodes in the family group that do not appear in the friends or work group.

More formally, to maximize utility, we want to maximize the number of nodes in F to which information is directly shared by the user.

6.5.3 Reconciling Utility and Privacy

Our goal, in this work, is to reconcile privacy goals, which try to bound the influence within the malicious node set M, with utility requirements, which try to maximize the number of nodes in F with which information is initially shared. As we saw above, both of these objectives are inherently at odds: for instance, increasing utility will necessarily also increase the influence within M. In order to reconcile privacy and utility, and thereby satisfy both requirements, we thus propose combined privacy and utility policies that optimize one criterion under a constraint on the other.

6.5.4 Policies for Information Propagation

As discussed above, we can approach reconciling privacy and utility requirements in diffusion network in two different ways. First, we can minimize the influence into the malicious node set while maintaining a lower bound on the number of friendly nodes we directly share information with. We formalize this in *utility-restricted privacy policies*.

Definition 28 (Utility-restricted Privacy Policy)**.** *A* utility-restricted privacy policy Π *is a 4-tuple* $\Pi = (F, M, k, t)$ *where* F *is the set of all friendly nodes,* M *is the set of all adversarial nodes,* k *is the number of friendly nodes the information should be shared to, and* t *is the period of time in which this policy should be valid.*

Second, instead of maximizing privacy while maintaining a given utility restriction, we maximize utility subject to a given privacy constraint. With a *privacy-restricted utility policy*, we maximize utility while adhering to a transmission (or privacy) threshold τ given as an upper bound on the expected number of $m \in M$ that receive the shared information within a time threshold t.

Definition 29 (Privacy-restricted Utility Policy)**.** *A privacy-restricted utility policy* Υ *is a 4-tuple* $\Upsilon = (F, M, \tau, t)$ *where* F *is the set of all friendly nodes,* M *is the set of all*

adversarial nodes, τ is the transmission threshold into M, and t is the period of time in which this policy should be valid.

Given such *propagation policies*, our goal is to find a subset $F' \subseteq F$ with which we initially share information and maximize utility and privacy under the constraints set by the policy. In the following sections we formally define what it means to satisfy the constraints set by a propagation policy for a given diffusion network as well as the tractability of optimally satisfying them.

6.5.5 Dropping the Time Threshold

Dropping the time threshold t in the above policies would be equivalent to setting $t = \infty$. In this case, for the continuous-time diffusion model, the problem of optimally satisfying a propagation policy typically reduces to finding a subset $F' \subseteq F$ that is part of a connected component of the network with minimal intersection with M: for $t \to \infty$, we get that $S_j(t \mid t_i, \alpha_{i,j}) \to 0$ if $\alpha_{i,j} > 0$ for diffusion networks with exponential, power law or Rayleigh transmission likelihoods [43]. Hence, any node reachable by F' is also infected by the shared information. In terms of complexity, we do not improve the tractability of optimally satisfying propagation policies by only considering $t \to \infty$, since we use this special case to show the NP-hardness of optimally satisfying these policies.

6.5.6 Privacy Guarantees and Information Propagation Policies

Clearly, the presented approach to controlling information propagation will not be able to provide hard security guarantees. The special case with non-finite time threshold discussed above presents important worst case: due to the high connectedness of today's social networks, essentially all nodes are reachable from each other. The above defined policies can therefore not be satisfied meaningfully for a non-finite time threshold t.

Even for finite time thresholds, our approach only provides a statistical risk estimate through the expected number of malicious nodes that will receive the shared information. Essentially, we try to minimize the *exposure* of a user's information to malicious nodes in the network, a concept explored by Mondal et al. [78] as an alternative to traditional access control principles: simply, in social network where a user can only control their own actions, they lose the ability to enforce any access control policies on their own information as soon as they share this information with their peers. Instead, the expected exposure of a piece of information can be used to regulate the amount of nodes in the network that this information reaches.

While controlling exposure is not robust against adversaries that actively look for information shared by the user, empirical evaluations by Mondal et al. show that users indeed care about the exposure of their information. The mechanism to minimize the exposure presented in this paper can therefore constitute a meaningful tool that enables users to control their privacy in a situation where traditional access control policies are not enforceable and provable privacy guarantees cannot meaningfully be provided.

6.6 Evaluating Information Propagation Policies

As discussed in the previous section, we want to achieve minimum influence into the malicious node set. We formalize this quantity under the notion of *exposure* that we need to minimize in order to optimally satisfy our policies. In the following we first define the notion of exposure and then derive the optimization problems that correspond to optimally satisfying utility-restricted privacy policies as well as privacy-restricted utility policies.

6.6.1 Exposure

In both, utility and privacy policies, we are interested in bounding the likelihood that an adversarial node gets infected by time t given that we initially infected a subset $F' \subseteq F$.

As introduced in Section 6.4, the influence $\sigma(A;t)$ determines the expected number of nodes infected by information initially shared with the set A. In our case, the initially infected set of nodes is F'. However, instead of determining the expected number of infected nodes in the whole network, we are now only interested in the expected number of infected nodes in the set of malicious nodes M. To this end, we define the notion of *exposure* that corresponds to exactly this quantity.

Definition 30 (Exposure)**.** *Let $N = (V, \alpha)$ be an information diffusion network. The* exposure $\chi_N(F, M, t)$ *caused by $F \subseteq V$ with respect to $M \subseteq V$ within time t is given by*

$$\chi_N(F, M, t) = \sum_{m_i \in M} \Pr[t_i \leq t \mid F].$$

Here, $\Pr[t_i \leq t \mid F]$ is the likelihood that the infection time t_i of malicious node m_i is smaller than t, given that F is infected at time 0.

We drop the subscript N if it is clear which network we consider from the context. Furthermore, we will write $\chi(F) = \chi(F, M, t)$ if both, the malicious node set M and time threshold t, are clear from the context as well.

Since our exposure function $\chi_N(F, M, t)$ essentially generalizes the regular influence function $\sigma_N(F, t)$ (cf. Section 6.4), computing $\chi_N(F, M, t)$ exactly is also $\#P$-hard. However, we can directly use the randomized approximation algorithm proposed by Gomez-Rodriguez et al. [46] for the influence function to approximate our exposure function up to a constant factor: we simply ignore the infection times for nodes not in M. In the following we will assume to have an oracle that exactly computes the exposure function for a given initial infection F.

6.6.2 Maximal Satisfaction of Propagation Policies

Equipped with the above exposure function, we can now define what it means that an initial infection of $F' \subseteq F$ within a network satisfies a utility-restricted privacy policy.

Definition 31. *An initial infection F' satisfies a utility-preserving privacy policy $\Pi = (F, M, \tau, t)$ in an information diffusion network N with an exposure τ if $F' \subseteq F$, $|F'| \geq k$ and $\chi(F', M, t) \leq \tau$.*

The set F' maximally satisfies Π in N if there is no other set $F'' \subseteq F$ with $|F''| \geq k$ and $\chi(F'', M, t) < \chi(F', M, t)$.

Similarly, we can also define when an initial infection F' satisfies a privacy-restricted utility policy.

Definition 32. *An initial infection of F' satisfies a privacy-preserving utility policy $\Upsilon = (F, M, \tau, t)$ in an information diffusion network N if $F' \subseteq F$ and $\chi(F', M, t) \leq \tau$.*

A set $F' \subseteq F$ maximally satisfies Υ in N if there is no other set $F'' \subseteq F$ with $|F''| > |F'|$ and $\chi(F'', M, t) \leq \tau$.

In order to maximize utility and privacy, we naturally want to find a maximally satisfying set of friends with which we initially share the information. In the following, we will see that this corresponds to solving constrained optimization problems.

6.6.3 Checking Policies as an Optimization Problem

To maximally satisfy a utility-restricted privacy policy, we need to find an initial infection of size k that minimizes the exposure. This yields the MAXIMUM-k-PRIVACY problem that we formalize below.

Definition 33 (MAXIMUM-k-PRIVACY)**.** *Given a utility-restricted privacy policy $\Pi = (F, M, k, t)$ and an information diffusion network N, the* MAXIMUM-k-PRIVACY *problem is given by*

$$\min_{F' \in 2^F} \chi(F', M, t) \text{ subject to } \left|F'\right| \geq k$$

Naturally, a solution to MAXIMUM-k-PRIVACY maximally satisfies the corresponding utility-restricted privacy policy.

Corollary 6. *If F' is an optimal solution to* MAXIMUM-k-PRIVACY *with respect to Π, then F' maximally satisfies Π.*

To maximally satisfy privacy-restricted utility policies, on the other hand, we want to find an initial infection of maximum size that results in an exposure below the privacy threshold τ. This yields the MAXIMUM-τ-UTILITY problem.

Definition 34 (MAXIMUM-τ-UTILITY)**.** *Given a privacy-restricted utility policy $\Upsilon = (F, M, \tau, t)$ and an information diffusion network N,* MAXIMUM-τ-UTILITY *problem is given by*

$$\max_{F' \in 2^F} \left|F'\right| \text{ subject to } \chi(F', M, t) \leq \tau$$

As for the utility-restricted privacy policy, a solution to the above optimization problem maximally satisfies the corresponding privacy-restricted utility policy.

Corollary 7. *If F' is an optimal solution to* MAXIMUM-τ-UTILITY *with respect to Υ, then F' maximally satisfies Υ.*

6.6.4 Checking Propagation Policies in Practice

Given the above optimization problems, a tool for controlling information propagation can be implemented in a straightforward manner: together with every action they perform, the user supplies either a utility-restricted privacy policy or a privacy-restricted utility policy. The tool then solves the corresponding optimization problem and return the subset $F^* \subseteq F$ with which the user can share the information to maximally satisfy the supplied policy.

Since these policies can differ for each action performed by the user, the tractability of the corresponding optimization problems is a major issue: in practice, the evaluation of the policies should happen live while the action is performed and should only cause minimal delay to ensure usability. In the following sections we therefore take a look at the tractability of the two optimization problems introduced above. While we will see that exactly solving these optimization problems is NP-hard, we will show that it can be approximated to a constant factor that depends on the problem instance.

6.7 Maximum-k-Privacy

In the following we take a look at the tractability of MAXIMUM-k-PRIVACY. Unfortunately, we will see that it is NP-hard and therefore does not allow for an efficient, exact solution. In a second step we will show that the objective function in MAXIMUM-k-PRIVACY is submodular with non-zero curvature. This allows us to at least efficiently approximate MAXIMUM-k-PRIVACY.

6.7.1 Hardness of Maximum-k-Privacy

To show that deciding MAXIMUM-k-PRIVACY is NP-hard, we give an efficient reduction of the MINIMUM k-UNION problem to MAXIMUM-k-PRIVACY. The MINIMUM k-UNION problem is defined as follows: given a collection of sets $\mathcal{C} = \{C_1, \ldots, C_j\}$ over a universe $\mathcal{U} = \bigcup_{C_i \in \mathcal{C}} C_i$ with $|\mathcal{U}| = n$ and a positive integer k with $k \leq j \leq n$, we want to find a subset $\mathcal{C}' \subseteq \mathcal{C}$ with $|\mathcal{C}'| \geq k$ that minimizes the cardinality of its union, i.e. $|\bigcup_{C_i \in \mathcal{C}'} C_i|$. MINIMUM k-UNION has been shown to be NP-hard by Vinterbo [106].

Theorem 3. MAXIMUM-k-PRIVACY *is NP-hard.*

Proof. Let $\mathcal{C} = \{C_1, \ldots, C_j\}$, $\mathcal{U}$ and $k \leq j \leq n$ be an instance of MINIMUM k-UNION. We construct a corresponding instance of MAXIMUM-k-PRIVACY as follows: We define our diffusion network $N = (V = \mathcal{C} \cup \mathcal{U}, \alpha)$ with transmission rate $\alpha_{i,j} = 1$ if $v_i \in \mathcal{C}$, $v_j \in \mathcal{U}$ and $v_j \in v_i$, and $\alpha_{i,j} = 0$ otherwise.

We now set the friendly node set as $F = \mathcal{C}$ and the malicious node set as $M = \mathcal{U}$. Deciding whether the original MINIMUM k-UNION instance has a solution of size l is then equivalent to deciding whether there is a subset $F' \subseteq F$ of size k with exposure $\chi(F', M, \infty) = l$, which is exactly MAXIMUM-k-PRIVACY with time-threshold ∞. □

6.7.2 Submodularity of Maximum-k-Privacy

We introduced the notion of submodularity in Section 6.4. In the following we show that our objective function $\chi(.)$ is submodular. Our proof follows the original proof of

submodularity for the general influence function in the continuous-time diffusion model presented by Gomez-Rodriguez and Schölkopf [45].

Theorem 4. *Given a network $N = (V, \alpha)$, two sets of nodes $F \in V$, $M \in V$ and a time horizon t, the exposure function $\chi_N(F, M, t)$ is a submodular function in the set of nodes F.*

Proof. We follow the proof of Theorem 4 in [45]. By definition, all nodes in F are infected at time $t_0 = 0$. Let Δt be some point in the probability space of the differences in infection times between each pair of nodes in the network. We define $\chi_{\Delta t}(F, M, t)$ as the total number of nodes in M infected in time less then t for Δt.

Define $R_{\Delta t}(f, M, t)$ as the set of nodes in M that can be reached from f in time shorter then t. Then $\chi_{\Delta t}(F, M, t) = \left|\bigcup_{f \in F} R_{\Delta t}(f, M, t)\right|$. Define $R_{\Delta t}(f|B, M, t)$ as a set of nodes in M that can be reached from node f in a time shorter then t, but cannot be reached from any node in the set of nodes $B \subseteq V$. It follows, that for any sets of nodes $B \subseteq B'$ it holds that $\left|R_{\Delta t}(f|B, M, t)\right| \geq \left|R_{\Delta t}(f|B', M, t)\right|$.

Consider sets of nodes $B \subseteq B' \subseteq V$, and a node $b \notin B'$. Following the diminishing returns property of submodular set functions, we get

$$\begin{aligned}
&\chi_{\Delta t}(B \cup \{b\}, M, t) - \chi_{\Delta t}(B, M, t) \\
=&\left|R_{\Delta t}(b|B, M, t)\right| \\
\geq&\left|R_{\Delta t}(b|B', M, t)\right| \\
=&\chi_{\Delta t}(B' \cup \{b\}, M, t) - \chi_{\Delta t}(B', M, t),
\end{aligned}$$

and thus $\chi_{\Delta t}(B, M, t)$ is submodular (for any Δt). Since $\chi(B, M, t)$ is the expectation of $\chi_{\Delta t}(B, M, t)$ over all Δt, and since submodularity is preserved under non-negative, weighted combinations, $\chi(B, M, t)$ is also submodular. □

In the following we use the submodularity of χ_N to derive an efficient constant factor approximation where the approximation factor depends on the structure of the underlying network N.

6.7.3 Approximating Maximum-k-Privacy

Svitkina and Fleischer show that submodular minimization under cardinality constraints can, in the worst-case, only be approximated up to a polynomial factor [102]. Iyer et al., however, show that this worst-case can be improved upon if our submodular objective function is monotone and has a non-zero *curvature* [53]. Let $f : 2^V \to \mathbb{R}$ be a monotone submodular function and let $f(j|V')$ for a subset $V' \subseteq V$ and $j \in V$ with $j \notin V'$ be defined as

$$f(j|V') = f(V' \cup \{j\}) - f(V').$$

Then the *curvature* κ_f of f is given by[1]

$$\kappa_f = \min_{j \in V} \frac{f\left(j|V \setminus \{j\}\right)}{f(\{j\})}$$

[1]Note that the definition of curvature presented here differs slightly from the one in [53] to save on some additive inversions. The results still carry over.

The worst-case approximation lower bound is only achieved for submodular function with a curvature of zero. If $\kappa_f > 0$ then the approximation algorithm proposed by Iyer et al. provides an approximation bound of $\frac{1}{\kappa_f}$. That is, let F' be the optimal solution found by their approximation algorithm for MAXIMUM-k-PRIVACY, and let F^* be optimal solution. Then

$$\chi(F', M, t) \leq \frac{1}{\kappa_f} \cdot \chi(F^*, M, t).$$

Since we only allow positive transmission rates $\alpha_{i,j}$ in our diffusion networks, χ is clearly monotone. The curvature of χ, however, depends heavily on the specific network N it is computed on. If we only consider finite time thresholds $t < \infty$, and as long as the survival function S is non-zero for all finite time frames and for all nodes in the network, we can show that the curvature of χ is also non-zero. Note that this condition is true for all typically considered diffusion models, including exponential, power law and Rayleigh models [43].

Theorem 5. *Let $t < \infty$ and Let $N = (V, \alpha)$ be a continuous-time diffusion network for which $S_j(t \mid t_i, \alpha_{i,j}) > 0$ for all $v_i, v_j \in V$. Then $\kappa_{\chi(F,M,t)} > 0$.*

Proof. First, note that we can simply remove all nodes v_i with $\chi(\{v_i\}, M, t) = 0$ from the friend set F since they do not generate any influence and can thus safely be infected at the beginning.

Now, let $v \in F$ be the node that minimizes $\kappa_{\chi(F,M,t)}$ and that satisfies the above condition. Clearly, $\chi(\{v\}, M, t) \neq 0$ if and only if there is a directed path from v to some node in M. Let $P = \{v_1, \ldots, v_m\}$ be the nodes on this path, with $v_m \in M$. Since we consider a finite time threshold t, and since the survival likelihood is non-zero for all pairs of nodes on this path, the cumulative infection likelihood for every node on this path is decreased if this path is removed. Thus, $\chi\left(v \mid F \setminus \{v\}, M, t\right) > 0$. □

The exact value of $\kappa_{\chi(F,\{m\},t)}$ depends heavily on the structure of the diffusion network. To exemplify this, we, in the following, give a lower bound for κ_χ on an example diffusion network. We consider a bipartite network with node sets F and M to simplify the computation of χ, and assume exponential transmission likelihoods and a bounded time threshold to compute an exact value for the curvature.

Lemma 2. *Let $N = (V = F \cup M, \alpha)$ be an exponential diffusion network with transmission rates $\alpha_{i,j} > 0$ only if $v_i \in F$ and $v_j \in M$, $\forall v_i, v_j \in V : \alpha_{i,j} \leq \alpha_{max}$ and maximum in-degree d. Then $\kappa_{\chi(F,M,t)} \geq e^{-\alpha_{max} \cdot d - 1}$ for $t = 1$.*

Proof. The exposure $\chi(F, M, t)$ for singleton malicious node set $M = \{m\}$ corresponds exactly to the likelihood that m is infected by time t by an initially infected set F. This infection likelihood is given by

$$T_m(t, F) = 1 - \prod_{f \in F \cap N(m)} S_m(t \mid \alpha_{f,m})$$

where $N(m)$ is the neighborhood of m and $S_m(t \mid \alpha_{f,m})$ is the likelihood that m survives without being infected from node f within time t (cf. Section 6.4). In the following we

will denote $F \cap N(m)$ with $F(m)$. Since each malicious node is infected independently, we can thus re-write our exposure function as

$$\chi(F, M, t) = \sum_{m \in M} T_m(t, F).$$

For the curvature $\kappa_{\chi(F,\{m\},t)}$ we thus get

$$\begin{aligned}\kappa_{\chi(F,\{m\},t)} &= \min_{x \in F} \frac{\sum_{m \in M} T_m(t, F) - T_m(t, F \setminus \{x\})}{\sum_{m \in M} T_m(t, \{x\})} \\ &= \min_{x \in F} \frac{\sum_{m \in M} \prod_{f \in F \setminus \{x\}(m)} S_m(t \mid \alpha_{f,m}) - \prod_{f \in F(m)} S_m(t \mid \alpha_{f,m})}{\sum_{m \in M} 1 - S_m(t \mid \alpha_{x,m})}\end{aligned}$$

Here, $\prod_{f \in F \setminus \{x\}(m)} S_m(t \mid \alpha_{f,m}) - \prod_{f \in F(m)} S_m(t \mid \alpha_{f,m})$ is zero if x is not in the neighborhood of m. We thus get

$$\begin{aligned}&= \min_{x \in F} \frac{(1 - S_m(t \mid \alpha_{x,m})) \sum_{m \in M(x)} \prod_{f \in F \setminus \{x\}(m)} S_m(t \mid \alpha_{f,m})}{\sum_{m \in M(x)} 1 - S_m(t \mid \alpha_{x,m})} \\ &= \min_{x \in F} \sum_{m \in M(x)} \prod_{f \in F \setminus \{x\}(m)} S_m(t \mid \alpha_{f,m}) \cdot \frac{1}{|M(x)|}\end{aligned}$$

In an exponential model, $S_m(t \mid \alpha_{x,m})) = e^{-\alpha_{x,m} t}$ [43]. We thus get

$$\min_{x \in F} \sum_{m \in M(x)} \prod_{f \in F \setminus \{x\}(m)} e^{-\alpha_{f,m} t} \cdot \frac{1}{|M(x)|}$$

which is minimized when all node in F are adjacent for the malicious nodes. Together with the upper bound on the transmission rates α_{max} and the time threshold $t = 1$, we get

$$\begin{aligned}&\min_{x \in F} \sum_{m \in M(x)} \prod_{f \in F \setminus \{x\}(m)} e^{-\alpha_{f,m} t} \cdot \frac{1}{|M(x)|} \\ \geq &|M(x)| \cdot e^{-\alpha_{max}(|F|-1)} \cdot \frac{1}{|M(x)|} \\ = &e^{-\alpha_{max}(|F|-1)}\end{aligned}$$

Given the in-degree bound d, we thus get $e^{-\alpha_{max}(d-1)}$ as the lower bound of the curvature (since x is always adjacent). □

With the non-zero curvature, we can apply the majorization-minimization algorithm proposed by Iyer et al. [53] to efficiently approximate MAXIMUM-k-PRIVACY upto to a factor of $\frac{1}{\kappa_\chi}$.

Theorem 6. *There is an efficient algorithm A that approximates* MAXIMUM-k-PRIVACY *to a factor $\frac{1}{\kappa_\chi}$. That is, let F' be the output of A and let F^* be the optimal solution. Then*

$$\chi(F', M, t) \leq \frac{1}{\kappa_\chi}\chi(F^*, M, t)$$

In the next subsection, we briefly recall the majorization-minimization algorithm specifically for our use-case and derive a complexity bound for the approximation.

6.7.4 The Majorization-minimization Algorithm

The constant factor approximation we derived above relies on the majorization-minimization algorithm put forward by Iyer et al. [53]. For the sake of being self-contained, we briefly recall this algorithm instantiated for our use-case and provide short complexity analysis.

The majorization-minimization algorithm begins with an arbitrary candidate solution Y to our optimization problem and then minimizes a modular approximation of the objective function along the *supergradients* of the objective function.

Definition 35. *The* supergradient g_Y *of the exposure function χ at a candidate solution $Y \subseteq F$ is given by*

$$g_Y(f) = \begin{cases} \chi(f \mid F \setminus \{f\}), & \text{if } f \in Y \\ \chi(f \mid Y), & \text{if } f \notin Y. \end{cases}$$

where $\chi(f \mid V)$ is given by

$$\chi(f \mid V) = \chi(V \cup \{f\}, M, t) - f(V, M, t).$$

The supergradient $g_Y(X)$ of a set of nodes $X \subseteq F$ is given by $g_Y(X) = \sum_{f \in X} g_Y(f)$.

Note that in their paper, Iyer et al. present a more general definition of supergradient that can be instantiated in various forms. For our use-case, however, the definition presented above is sufficient to achieve the required approximation guarantees.

To improve on the candidate solution, the majorization-minimization algorithm minimizes the modular approximation of our exposure function at the candidate solution.

Definition 36. *The* modular approximation m *of the exposure function χ at a candidate solution $Y \subseteq F$ is given by*

$$m^{g_Y}(X) = \chi(Y) + g_Y(X) - g_Y(Y).$$

Due to the submodularity of our exposure function χ we can use this modular approximation as an upper bound for χ, i.e. $m^{g_Y}(X) \geq \chi(X)$ [53].

An adaption of the majorization-minimization approximation algorithm to our use-case is illustrated in Algorithm 2. We can easily see that each step improves the found

Algorithm 2 Approximating Maximum-k-Privacy

Require: Instance F, M, k of MAXIMUM-k-PRIVACY
1: $\mathcal{C} \leftarrow \{X \subseteq F \mid |X| = k\}$
2: Select random candidate solution $X^1 \in \mathcal{C}$
3: $t \leftarrow 0$
4: **repeat**
5: $\quad t \leftarrow t + 1$
6: $\quad X^{t+1} \leftarrow \arg\min_{X \in \mathcal{C}} m^{g_{X^t}}(X)$
7: **until** $X^{t+1} = X^t$
8: **return** X^t

Algorithm 3 Minimizing g_Y

Require: Instance F, M, k of MAXIMUM-k-PRIVACY, $\forall f \in F : \chi(f \mid F \setminus \{f\})$
1: $\forall f \in Y$: compute $\chi(f \mid Y)$
2: Let Π_F be F ordered by $g_Y(f)$ in ascending order
3: $X \leftarrow$ first k elements of Π_F
4: **return** X

solution: in each iteration, the new solution X^{t+1} minimizes the modular approximation, and thus

$$\chi(X^{t+1}) \leq m^{g_{X^t}}(X^{t+1}) \leq m^{g_{X^t}}(X^t) = \chi(X^{t+1}).$$

The approximation bound derived in the previous subsection is already achieved after the first iteration. The complexity of this one iteration is essentially the minimization performed in step 6. This minimization can be done exactly in an efficient manner: by definition, minimizing $m^{g_Y}(X)$ is equivalent to minimizing $g_Y(X)$. We can minimize the supergradient $g_Y(X)$ using the greedy approach in time linear in $|F|$.

As a simplification, we can pre-compute the values $\chi(f \mid F \setminus \{f\})$ for all $f \in F$ since these values are independent of the candidate solution Y. During the minimization of g_Y, we then only need to compute $\chi(f \mid Y)$ for all $f \in Y$ before we then sort all values $g_Y(f)$ and apply the greedy mechanism. A pseudocode implementation of the resulting algorithm for minimizing $g_Y(X)$ is given in Algorithm 3.

Assuming we need R computation steps to perform one exposure computation, and since g_Y is real valued and can thus be sorted in linear time using Radix-Sort or a similar sorting algorithms, an upper bound for the running time of one iteration of Algorithm 2 is given by $O(k \cdot R + |F|)$. The step of pre-computing $\chi(f \mid F \setminus \{f\})$ for all $f \in F$ takes $O(|F| \cdot R)$ time.

6.8 Maximum-τ-Utility

In the following, we take a look at MAXIMUM-τ-UTILITY and its tractability. We first show that MAXIMUM-τ-UTILITY is NP-hard as well and therefore likely does not allow for an efficient, exact algorithm. Since MAXIMUM-τ-UTILITY and MAXIMUM-k-PRIVACY are closely related, however, we are able to leverage the approximation

algorithm for MAXIMUM-k-PRIVACY discussed in the previous Section to provide a constant factor approximation for MAXIMUM-τ-UTILITY as well.

6.8.1 Hardness of Maximum-τ-Utility

To show the NP-hardness of MAXIMUM-τ-UTILITY, we first consider the MAXIMUM l-UNION problem defined as follows: given a collection $\mathcal{C} = C_1, \ldots, C_j$ over a universe $\mathcal{U} = \bigcup_{C_i \in \mathcal{C}} C_i$ with $|\mathcal{U}| = n$ and a parameter l with $l \leq |\mathcal{U}|$, find the largest subset $\mathcal{C}' \subseteq \mathcal{C}$ with $|\bigcup_{C_i \in \mathcal{C}'} C_i| \leq l$. MAXIMUM l-UNION is NP-hard by reduction from MINIMUM k-UNION, which we introduced in the previous section.

Lemma 3. MAXIMUM l-UNION *is NP-hard.*

Proof. Let $\mathcal{C} = C_1, \ldots, C_j, \mathcal{U}$ and $k \leq j \leq n$ be an instance of MINIMUM k-UNION. We now perform a binary search for l^* in $[0, n]$ such that MAXIMUM l^*-UNION has a solution of size k, but MAXIMUM $l^* - 1$-UNION has solution of size $k' < k$. The union size l^* is then the minimum size of a k-union over $\mathcal{C}$, and thus the solution to our MINIMUM k-UNION instance. □

Using the hardness of MAXIMUM l-UNION, we can now also show the hardness of MAXIMUM-τ-UTILITY.

Theorem 7. MAXIMUM-τ-UTILITY *is NP-hard.*

Proof. Given an instance $\mathcal{C} = C_1, \ldots, C_j$, $\mathcal{U}$ and l for MAXIMUM l-UNION, we perform the same reduction as in the proof of Theorem 3. On the resulting network N, MAXIMUM-τ-UTILITY with an influence threshold of l has a solution of size k if and only if the instance for MAXIMUM l-UNION has a solution of size k. Since by Lemma 3, MAXIMUM l-UNION is NP-hard, MAXIMUM-τ-UTILITY is also NP-hard. □

6.8.2 Approximation of Maximum-τ-Utility

Taking a close look at Definition 34 and Definition 33, we can derive a straightforward algorithm to solve MAXIMUM-τ-UTILITY given an algorithm for MAXIMUM-k-PRIVACY: we iterate over the values of k starting with $k = |F|$ down to 1 and compute the minimum exposure subset $F' \subseteq F$ of size k. If for some k we find a subset F' with $\chi_N(F', M, t) \leq \tau$, we stop and return F'. A pseudocode implementation of this algorithm is shown in Algorithm 4.

Algorithm 4 Optimizing Maximum-τ-Utility

Require: Instance F, M, τ of MAXIMUM-τ-UTILITY
1: **for** $n \in [|F|, \ldots, 1]$ **do**
2: $\quad \tau' := \min_{F' \subseteq F} \chi(F')$ s.t. $|F'| = n$
3: $\quad$ **if** $\tau' \leq \tau$ **then return** n
4: $\quad$ **end if**
5: **end for**
6: **return** 0

Naturally, if we have an efficient and exact algorithm to solve MAXIMUM-k-PRIVACY, the above algorithm provides an exact solution to MAXIMUM-τ-UTILITY. However, as discussed in Section 6.7, we are only able to efficiently approximate MAXIMUM-k-PRIVACY up to a factor of $\frac{1}{\kappa_\chi}$. Still, using this approximation algorithm in Algorithm 4, we overall obtain an κ_χ-approximation for MAXIMUM-τ-UTILITY.

Theorem 8. *Let n^* be the optimal solution to an instance of* MAXIMUM-τ-UTILITY*, and let n be the output of Algorithm 4 to the same instance, using an $\frac{1}{\kappa_\chi}$-approximation for* MAXIMUM-k-PRIVACY*. Then $n \geq \kappa_\chi n^*$.*

Proof. We first prove the following claims that we will require afterwards for the proof of above statement.

Claim 1.

$$\begin{aligned}\forall F' \subseteq F : \chi(F) &\geq \chi(F') + \kappa_\chi \sum_{f \in F \setminus F'} \chi(f) \\ &\geq \chi(F') + \kappa_\chi(|F| - |F'|)\chi_{min}\end{aligned}$$

where $\chi(f)$ is short for $\chi(\{f\})$ and $\chi_{min} = \min_{f \in F} \chi(f)$.

Proof. By the definition of κ_χ, it holds that

$$\chi(F \setminus \{f\}) \leq \chi(F) - \kappa_\chi \chi(f).$$

Iteratively applying this inequality, we get that

$$\begin{aligned}\forall F' \subseteq F : &\chi(F') \leq \chi(F) - \kappa_\chi \sum_{f \in F \setminus F'} \chi(f) \\ \Leftrightarrow &\chi(F) \geq \chi(F') + \kappa_\chi \sum_{f \in F \setminus F'} \chi(f)\end{aligned}$$

Now, since $\forall f \in F : \chi_{min} \leq \chi(f)$, it also holds that

$$\chi(F) \geq \chi(F') + \kappa_\chi(|F| - |F'|)\chi_{min}$$

As a special case, since $\chi(\emptyset) = 0$, we get that

$$\chi(F) \geq \kappa_\chi \sum_{f \in F} \chi(f) \geq \kappa_\chi |F| \chi_{min}.$$

□

In the following, we denote

$$\chi(n) = \min_{F' \in 2^F} \chi(F') \text{ } s.t. \text{ } |F'| = n.$$

Claim 2. *For all $F' \subseteq F$ and integer $n \leq |F'|$ it holds that*

$$\chi(F') - \chi(n) \geq \kappa_\chi(|F'| - n)\chi_{min}.$$

Proof. By Claim 1, we have that

$$\forall F'' \subseteq F' : \chi(F') \geq \chi(F'') + \kappa_\chi(|F'| - |F''|)\chi_{min}.$$

Since $\chi(n) \leq \chi(F'')$ for all subsets F'' with $|F''| = n$, we get that

$$\chi(F') \geq \chi(n) + \kappa_\chi(|F'| - n)\chi_{min}.$$

By re-arranging, we get our claim. □

We now prove the statement of Theorem 8. Let $\chi'(n)$ be the output of the approximation algorithm for MAXIMUM-k-PRIVACY with size-constraint n, and let $\chi(n)$ be the optimal solution. We are thus looking for an approximation factor x with $\chi'(x \cdot n) \leq \chi(n)$. Since our approximation algorithm for MAXIMUM-k-PRIVACY has an approximation factor of $\frac{1}{\kappa_\chi}$, we have that

$$\chi'(x \cdot n) \leq \frac{1}{\kappa_\chi}\chi(x \cdot n)$$

With Claim 2, we then get

$$\begin{aligned} &\leq \frac{1}{\kappa_\chi}(\chi(n) - \kappa_\chi(1-x) \cdot n \cdot \chi_{min}) \\ &\leq \chi(n) + (\frac{1}{\kappa_\chi} - 1)\chi(n) - (1-x) \cdot n \cdot \chi_{min} \end{aligned}$$

The last statement is $\leq \chi(n)$ only if

$$(\frac{1}{\kappa_\chi} - 1)\chi(n) - (1-x) \cdot n \cdot \chi_{min} \leq 0$$

Let $y = 1 - x$. We then require

$$\begin{aligned} y &\geq (\frac{1}{\kappa_\chi} - 1)\frac{\chi(n)}{n \cdot \chi_{min}} \\ &\geq (\frac{1}{\kappa_\chi} - 1)\frac{\kappa_\chi \cdot n \cdot \chi_{min}}{n \cdot \chi_{min}} && \text{(by Claim 1)} \\ &\geq 1 - \kappa_\rho \end{aligned}$$

and thus $x \leq \kappa_\rho$. Subsequently, for all $n' \leq \kappa_\rho n$ it holds that $\chi'(n') \leq \chi(n)$, and thus the theorem follows. □

With this results we thus showed that both, MAXIMUM-k-PRIVACY and MAXIMUM-τ-UTILITY, while NP-hard, can be efficiently approximated to a constant factor in the curvature of our exposure function. The majorization-minimization algorithm for constrained submodular minimization proposed by Iyers et al. [53] that we leverage for this result solves much more general minimization problems and it would be an interesting direction for future to see whether there is a more specialized algorithm with even better approximation guarantees.

6.9 Discussion

We finally discuss advantages and limitations of the approach presented in this paper. Furthermore, we also present potential directions for future work.

6.9.1 Limitations

The approach presented in this paper leverages diffusion networks to control information propagation while at the same time maximizing the user's utility. Since this approach is parametric in the network model, the quality of the exposure minimization inherently depends on the accuracy of the transmission parameters in the model. In particular, this approach also does not provide any provable bounds on information propagation risks unless the transmission rates in the diffusion model are provably correct.

6.9.2 Privacy and Utility Parameters

A general issue of various privacy mechanisms in practice is that their guarantees are very hard to understand, for layman as well as experts. The approach presented in this paper, however, provides utility and privacy bounds that are very easy to understand: utility directly corresponds to the number of friendly nodes with which we are allowed to share information, while our exposure function encodes the expected number of malicious nodes that is going to receive the shared information within a time frame t.

6.9.3 Finding Bad Apples

An alternative approach to the utility-constrained exposure minimization introduced in Section 6.6 as MAXIMUM-k-PRIVACY is to identify the node $f \in F$ that has the largest contribution to the exposure function $\chi(F, M, t)$. Since the contribution of each node v to χ is not constant, this is, however, not as easy as simply comparing $\chi(\{f\}, M, t)$ for each $f \in F$. In particular under utility constraints, we are more interested in the marginal contribution of f when added last to the initial infection. That is, we want to find

$$f = \arg\min_{f' \in F} \min_{F' \in 2^{F \setminus \{f'\}}} \chi(F', M, t) \text{ s.t. } |F'| = |F| - 1.$$

Taking a close look at MAXIMUM-k-PRIVACY, however, we quickly see that this is equivalent to finding the element $f \in F \setminus F^*$ where

$$F^* = \arg\min_{F' \in 2^F} \chi(F', M, t) \text{ s.t. } |F'| = |F| - 1.$$

Finding bad apples is therefore a special case of MAXIMUM-k-PRIVACY.

6.9.4 Potential Extension

We take a look at two natural extensions to our policy and exposure definitions that do not significantly alter the tractability results presented in this paper. We did not include them into our original definitions to keep them and our proofs simple.

The propagation policy definitions in Section 6.5.1 assume uniform, unit utility that is gained from sharing information with nodes within F. This can naturally be extended by also providing a utility function $\mu : F \to \mathbb{R}^+$ that defines the payout we receive for sharing information with a friendly node $f \in F$. For MAXIMUM-k-PRIVACY, this changes the cardinality constraint to a knapsack constraint. That is, we need to solve the following minimization problem

$$\min_{F' \in 2^F} \rho(F, M, t) \text{ s.t. } \sum_{f \in F'} \mu(f) \geq k$$

However, since the majorization-minimization approximation algorithm we leverage in Section 6.7 also works with such knapsack constraints, the tractability of approximating MAXIMUM-k-PRIVACY does not change.

The definition of the exposure function χ presented in Section 6.6.1 can be extended to include non-unit costs $c : M \to \mathbb{R}^+$ for each malicious node that is infected, similarly to the extension of the propagation policies with utility functions discussed above. The augmented exposure function would then be given by

$$\chi^{\mathrm{c}}_N(F, M, t) = \sum_{m_i \in M} c(m) \cdot \Pr[t_i \leq t \mid F].$$

The submodularity proof for simple exposure function naturally generalizes to this augmented exposure function: as long as the cost function is non-zero and positive, the diminishing return property still holds. Since, therefore, the augmented exposure function is submodular as well, we can directly adopt the results from the approximation of the simple exposure function. Note, however, that the curvature of the augmented exposure function will differ from the curvature of the simple exposure function.

6.9.5 Global vs. local view

In this paper, we implicitly assumed to have access to a diffusion model that correctly represent the diffusion properties of the whole network. Such a *global view* of the network, is, for instance, possible for the provider of a social networking service. Relying on a diffusion model provided by a service provider is, however, a) not always possible, or b) not always in the interest of the user. Therefore, the evaluation of propagation policies on an incomplete, *local view* of the network with provable bounds on the quality of the evaluation results would be desirable. Such a local view would replace parts of the actual transmission matrix with transmission rates $\hat{\alpha}_{i,j}$ drawn from a distribution $\mathcal{A}$.

Let $\hat{\alpha}$ denote the transmission matrix consisting of the partial transmission matrix α and the randomly chosen values $\hat{\alpha}_{i,j}$ for the unknown transmission rates. Let $\hat{N} = (V, \hat{\alpha})$. To soundly approximate the policy evaluation on the global view, we then need to solve the following optimization problem to maximally satisfy utility-restricted privacy policies

$$\min_{F' \in 2^F} \max_{\hat{N}} \chi_{\hat{N}}(F', M, t) \text{ subject to } |F'| \geq k.$$

However, because the exposure function χ increases monotonically in the transmission rates, the above optimization problem is equivalent to

$$\min_{F' \in 2^F} \chi_{N_{max}}(F', M, t) \text{ subject to } |F'| \geq k$$

where $N_{max} = (V, \alpha_{max})$ and α_{max} is the transmission matrix that uses the maximal value for all unknown transmission rates. The maximal satisfaction of propagation policies for the local view therefore has the same complexity and approximability as for the global view. The question of how to provide a general bound on the difference of the computed solution is, however, yet unanswered, and provides a direction for potential future work.

6.9.6 Hardness in Practice

As discussed in Section 6.6.4 the tractability of MAXIMUM-k-PRIVACY and MAXIMUM-τ-UTILITY is a critical issue if we want to leverage our approach for tools that control information propagation in practice. While we showed that both of these problems are NP-hard to solve exactly, in practice, one might be able to leverage advanced algorithmic techniques to augment the online optimization of the exposure function with (expensive) offline auxiliary computations that reduce the online computation time.

6.10 Conclusion

In this chapter, we tackled the challenge controlling information propagation in social networks while maintaining utility. We leverage diffusion networks to model the information propagation behavior in social networks and present two approaches to reconciling privacy and utility. First, the utility-restricted privacy policies and the corresponding MAXIMUM-k-PRIVACY optimization problem in which we minimize the likelihood of propagating information to malicious nodes under a lower bound constraint on the number of friendly nodes to which we initially share this information. And second, the privacy-restricted utility policies, and the corresponding MAXIMUM-τ-UTILITY optimization problem, in which we maximize the number of friendly nodes to which we initially share a piece of information subject to an upper bound constraint to the expected number of malicious nodes this information reaches.

We show that, while both optimization problems are NP-hard, the submodularity of our exposure function allows us to efficiently approximate them to a constant approximation factor that depends on the structure of the underlying diffusion network.

6.11 Future Work

For future work, a major open question is how much the optimization results computed for a local view of the propagation network differ from the exact optimization result. While we discussed a first step in this direction in Section 6.9.5, it is still unclear how to approach providing a general bound.

Another direction is to extend the diffusion models we consider in this work to also include node-dependent adaptation rates for different types of information. This could help in modeling in difference in interest of different users in a piece of information depending on its content.

7
XpoMin

Towards Practical Exposure Minimization in Continuous Time Diffusion Networks

7.1 Motivation

Sharing Information about their lives has become a daily routine for millions of user of various social media platforms such as Facebook and Twitter. Numerous cases have shown that the information shared on those platforms can easily and rapidly reach a very large audience – or even go viral – by means of sharing, re-tweeting, etc. This transitive diffusion of information leaves users with little to no control how and to whom which pieces of information will spread, oftentimes resulting in situations where information is disclosed to recipients that it was never supposed to reach. Users naturally have an interest in retaining control over their sensitive information and in limiting their diffusion (privacy).

If one takes the perspective of a privacy fundamentalist and strives for strong privacy guarantees as the superordinate goal, the only acceptable solution to this problem in open systems without trusted third parties that regulate the diffusion of information is to never disclose any information in the first place, since a powerful, actively probing adversary can easily learn any publicly shared information almost instantaneously. This worst-case perspective is of course rarely followed in practice since it contradicts the users' interest in sharing information with their intended recipients (utility). The goal is instead to empower the user to understand, assess, and where possible control the exposure of shared information while at the same time fulfilling their legitimate utility requirements [78].

7.2 Problem Description

In the previous chapter, we discussed the use of network diffusion models to mathematically study the propagation of information through a network of interacting users that share information [3, 43]. They have moreover been used to estimate the *influence* of a network node, i.e., the expected number of nodes that a certain piece of information will reach once it is shared by this node [46]. This measure has subsequently proved useful for determining network structures that maximize the achievable influence, with use cases in, e.g., viral marketing [17].

We then leveraged the concept of diffusion models to introduce the so-called Maximum-k-Privacy problem that formalizes the aforementioned privacy and utility requirements and their inherent tension. Formally, this problem pertains to finding a subset of *friendly nodes* of size k that contains a piece of information and which minimizes the *exposure*, i.e., the expected number of *malicious nodes* in the network that will receive the shared information within time T through the global diffusion behavior of the underlying network. The set of friendly and malicious nodes here are considered parameters of the problem instance. It has been shown that Maximum-k-Privacy is NP-hard, but that it can be efficiently approximated to the factor $\frac{1}{\kappa}$ where the *curvature* κ is a problem instance-specific constant.

While this provides a first theoretical underpinning of the problem under consideration, applying the approach to practical settings still entails formidable research challenges. First and foremost, the only existing approximation algorithm for Maximum-k-Privacy assumes oracle access to a solution of the exposure estimation problem

which is itself known to be $\#P$-hard. Any successful practical deployment of this algorithm would hence necessitate an efficient instantiation of this oracle that in particular satisfies critical timing requirements such as near-instant response times to queries is crucial. No such instantiation is known. Second, we still lack any insights in how the formal parameters correspond to real-data networks, e.g., how small the curvature κ is for realistic use cases, and how frequently the corresponding algorithmic worst cases appear in practice.

7.3 Contributions

Our contributions in this chapter are two-fold. First, we present XpoMin, a framework for the constrained exposure minimization in continuous time diffusion networks that, among other heuristics, implements the approximation algorithm for the Maximum-k-Privacy problem. To implement the efficient exposure estimation that also satisfies the aforementioned critical timing constraints we adapt a recently proposed, scalable approach to the regular influence estimation problem by Du et al. [28]. We generalize this algorithm to fit the exposure estimation problem that generalizes the influence estimation problem and prove the correctness of this generalization. The structure of this algorithm allows us to partition it into three distinct computation phases that, depending on the use case, can be pre-computed offline to build a data structure that subsequently allows for doubly logarithmic query times for the exposure estimation at the cost of linear memory requirements (both in the number of nodes in the network).

We evaluate the performance of our implementation in terms of memory usage as well as run time for the different computation phases. To this end we utilize a set of synthetic Kronecker networks that have been used to in the past to simulate diffusion behavior in social networks [67] as well as a real diffusion network based on the MemeTracker methodology [66]. In these evaluations, our prototypical implementation achieves a running time of under 800ms for the exposure minimization in the worst case, with overall constant computation times for the minimization across network sizes. Despite the substantial computations required for the pre-computation phases, this shows that, when offered as a service, exposure minimization can realistically be used in practice as an advisory tool for information sharing.

For our second contribution, we evaluate the accuracy of the exposure minimization performed by XpoMin. Using the same networks mentioned above, we first determine the value of the curvature κ (from which we can derive theoretical approximation bounds) for different Maximum-k-Privacy instances on these networks. We then perform the influence minimization with our approximation algorithms as well as with complete search (for smaller networks) and compare the outputs. Our evaluations show that the curvature itself can regularly reach very small values in the range from 10^{-4} to 10^{-1}, which leads to a corresponding worst case approximation guarantee that is insufficient in practice. The actual approximation results, however, are much closer to the optimal solution: on average, we achieve a worst case relative error $\delta < 2$ where δ is the ratio $\delta = \frac{\hat{\chi}^*}{\chi^*}$ between approximated minimum exposure $\hat{\chi}^*$ and actual optimal exposure χ^*.

Consequently, XpoMin paves the way for a practical solution to the exposure min-

imization problem. The implementation underlying XpoMin and the data used in the evaluations presented in this chapter will be made available publicly after the publication of corresponding paper [P4].

Outline In Section 7.5, we describe the main algorithms of XpoMin and their implementation 7.5. We then evaluate the performance of XpoMin, both in terms of memory and computation time, in Section 7.7, which concludes our first contribution. We then turn to our second contribution in Section 7.8, where we analyze the curvature of problem instances based on synthetic as well as real data and evaluate the actual approximation performance of the approximation algorithms implemented in XpoMin. In Section 7.9, we summarize our evaluation results as well as discuss various important aspects of applying XpoMin in practice. We finally conclude in Section 7.10.

7.4 Scalable Influence Estimation

We briefly recall the scalable influence estimation algorithm put forward by Du et al. [28] which we will later extend for the estimation of exposure in a diffusion network. The problem of estimating the influence caused by an initial infection A has been shown to be $\#P$-complete for the independent cascade model by Chen et al. [17], as well as for the continuous-time diffusion model by Gomez-Rodriguez et al. [46]. Recently, Du et al. proposed a scalable, sampling based approximation algorithm for the influence estimation problem [28]. The main idea behind their approach is as follows: instead of computing the likelihood of infection for every node in the network (which was done by previous, less efficient approaches to estimating influence [46]) they instead directly estimate the expected number of infected nodes. This overall allows for a highly parallelizable, and therefore scalable algorithm for estimating the influence in a diffusion network. In this paper, we will adapt their algorithm for our use-case of minimizing exposure.

Du et al.'s algorithm uses an important property of the exponential distribution[1]: given a collection of k random variables $X_i \sim \exp(1)$, the minimum X^* of these random variables is distributed according to $X^* \sim \exp(k)$. Accordingly, the minimum label assigned to any node reachable by A is distributed according to $l_j^* \sim \exp(|\mathcal{N}(A,t)|)$ where $|\mathcal{N}(A,t)|$ is the size of the neighborhood of nodes A within distance T. Using the m different labellings for a distance network sample, we can thus get the size of the neighborhood by estimating the parameter of the exponential distribution from which the smallest labels were drawn. The corresponding unbiased maximum likelihood estimator [20] is

$$|\mathcal{N}(A,T)| \approx \frac{m-1}{\sum_{j=1}^{m} l_j^*}.$$

Finally, they show that by averaging this neighborhood size over all n distance network samples of the network they obtain a constant factor approximation of the influence

[1] To simplify notation, we will henceforth denote with $\exp(a)$ the exponential distribution with mean a (or rate parameter $\frac{1}{a}$) with the probability density function $f(x) = \frac{1}{a}e^{-\frac{1}{a}x}$ if $x > 0$, and 0 otherwise.

Algorithm 5 (Malicious) Least Label List

Require: Reversed directed Graph $G = (V, E)$ with edge weights $\{\tau_{i,j}\}_{(i,j)\in E}$, node labels $\{l_i\}_{i\in V}$

for $i \in V$ **do**
  $d_i = \infty$
  $l^*(i) = \emptyset$
end for
for (malicious nodes) i sorted by l_i **do**
  heap $H = \emptyset$
  set all nodes as unvisited
  set i as visited
  push $(0, i)$ into H
  while $H \neq \emptyset$ **do**
    pop (d^*, s) with minimum d^* from H
    add (d^*, r_i) to $l^*(s)$
    $d_s = d^*$
    for unvisited neightbor j of s **do**
      set j as visited
      if $(d, j) \in H$ **then**
        Pop (d, j) from H
        Push $(\min(d, d^* + \tau_{j,s}), j)$ into H
      else
        if $d^* + \tau_{j,s} < d_j$ **then**
          Push $(d^* + \tau_{j,s}, j)$ into H
        end if
      end if
    end for
  end while
end for
return l^*

$\sigma_N(A; t)$ of the set A, where the constant factor depends on the number n of distance network samples drawn.

The above algorithm uses the least label list data structure to quickly find the smallest random label assigned to the neighborhood of A. To compute these least label lists Cohen proposes a modified Dijkstra's algorithm [20]. We present a pseudocode implementation in Algorithm 5. This algorithm produces, given labels l_i for each node v_i in a distance network, and distances $d_{i,j}$ between from nodes v_i to v_j, for each v in the network a list $l^*(v)$ of ordered pairs (d, l) with following property:

$$\begin{gathered} \infty > d_{(1)} \geq d_{(2)} \geq \ldots > d_{(|l^*(v)|)} \geq 0 \\ l_{(1)} \leq l_{(2)} \leq \ldots \leq l_{(|l^*(v)|)} \end{gathered} \tag{7.1}$$

In essence, Algorithm 5 works as follows: given a directed distance graph G, iterate over all nodes i in G ordered by their labels and perform a Breadth-First-Search in the

reversed distance graph G_{rev} where the direction of all edges is reversed. Doing this, you find the distance $d_{j,i}$ of the node i to all nodes j that can reach i in the original graph G, and append $(d_{j,i}, l_i)$ to j' least label list $l^*(j)$ if there has not been another node with smaller distance that was already added to $l^*(j)$. In consequence, following conditions hold at each iteration:

A) nodes are added to least label list in increasing order of their labels.

B) a node i is added to a least label list $l^*(j)$ if and only if all existing entries have a larger distance to j than i.

With these conditions, it is easy to see that Property 7.1 holds for the computed least label lists. The time complexity of Algorithm 5 is $O(|E| \log|V| + |V| \log^2|V|)$ [20].

To now determine the smallest label in the neighborhood of a node j within distance d, a simple binary search over $l^*(j)$ by the distance values is sufficient. Cohen shows that the expected size of each least label list is $O(\log|V|)$. Together with the binary search this provides a $O(\log\log|V|)$ algorithm to find the smallest label. Since we perform this search for $n \cdot m$ labeling of the graph for the influence estimation, this overall yields a $O(n \cdot m \cdot \log\log|V|)$ algorithm to estimate the influence of a single node. If we want to estimate the influence of a collection of nodes A, we simply have to find the smallest label within the neighborhoods of all nodes $v_i \in A$. Consequently, this leads to a $O(n \cdot m \cdot |A| \log\log|V|)$ algorithm for this case. However, the size of the initial infection $|A|$ is typically much smaller than $|V|$, thus the actual influence estimation remains fast.

7.5 XpoMin Methodology

In the following we develop the XpoMin framework. We first discuss the challenges of exposure minimization we face in practice. Then, we adapt the scalable influence estimation algorithm discussed above to our proble of estimating exposure in continuous time diffusion networks. We show that our adaptation correctly approximates the exposure function and provide an analysis of its time and space complexity. In a second step we then use this exposure estimation algorithm to implement the exposure minimization algorithm developed in the previous chapter. We also discuss alternative heuristics for exposure minimization that, while not as accurate, might allow for a faster exposure minimization.

7.5.1 Exposure Minimization in Practice

We implement the approximation algorithm for Maximum-k-Privacy we developed in Chapter 6 with particular focus on dealing with computation-time constraints typically faced in practice. This includes, for instance, very fast query response times during runtime at the cost of a long pre-computation phase in which the while diffusion network is processed. Since the notion of exposure generalizes influence, the sub-problem of estimating exposure is also at least $\#P$-hard since the influence estimation problem is also $\#P$-hard [17, 108]. This is especially problematic because the runtime of the

exposure minimization algorithm mentioned above directly depends on the runtime of the exposure estimation. To efficiently approximate the estimation of exposure, we therefore adopt the scalable influence estimation algorithm by Du et al. [28] for the notion of exposure.

As we will see, however, this still does not yield an algorithm that could effectively be computed by users themselves. While we will achieve near instant query times for the exposure minimization, the necessary pre-computations rely on A) knowing the whole diffusion network its diffusion parameters, and B) processing the whole diffusion network during the algorithm to obtain sound exposure estimates for the exposure minmization. Considering the limited computational capabilities that can typically be expected from an average user, these requirements make the implementation presented in this chapter, even as approximations, infeasible to be performed on user side.

Instead, we invision two applications for the approach presented in this chapter: first, it can be used by service providers to allow their users to perform the exposure minimization using their help. Service providers have access to the whole social network they operate, and can, in particular, easily collect the necessary data to infer diffusion paramters for the network. Furthermore, a large part of the pre-computations need to be performed only once for the whole network, and the exposure minimization algorithm is highly scalable, allowing for an effective computation by the service providers.

Secondly, our results can be used as a baseline for future work on user-side exposure minimization. In particular, the results of the XpoMin framework can be used as a benchmark for the performance of exposure minimization algorithms that work with, e.g., limited information or limited computations to overcome the user-side limitations mentioned above.

7.5.2 Scalable Exposure Estimation

To estimate the exposure $\chi(F, M, t)$ of an initial infection F with respect to a set of malicious nodes M, we only need to slightly modify Algorithm 5 (cf. Section 7.4): instead of iterating over all nodes in the network in line 4, we instead only iterate over the set of malicious nodes. Through that, we determine a *malicious least label list* l^*_m that only contains malicious nodes and still fulfills the conditions A and B mentioned above. Again, it directly follows that Property 7.1 also holds for the malicious least label list and therefore the modified least label list algorithm correctly produces the malicious least label lists.

By directly applying the complexity analysis performed by Cohen [20], we also obtain an expected length of $O(\log|M|)$ (see [20, Proposition 5.4]) for the malicious least label lists, a query time of $O(\log\log|M|)$, and a running time for the modified Algorithm 5 of $O(|E|\log|M| + |V|\log|V|\log|M|)$. Compared to the original least label list algorithm we only save a minor logarithmic factor if $|M| << |V|$.

Given the malicious least label lists, we can estimate the size *the malicious neighborhood*

$$\mathcal{M}(A, M, t) = \{v_j \in M \mid \exists v_i \in A : d_{i,j} \leq t\}$$

in the same way as the regular neighborhood size: the minimum of $|\mathcal{M}(A, M, t)|$ random variables drawn from exp(1) is distributed according to exp($|\mathcal{M}(A, M, t)|$). Con-

Algorithm 6 Exposure Estimation

Require: Diffusion Network $N = (V, \alpha)$, initial infection $A \subseteq V$, malicious nodes $M \subseteq V$, time threshold t

1: Sample n directed distance networks N_i
2: Sample m random labelings for each distance network N_i
3: **for** Each distance network sample N_i **do**
4: **if** $\exists v_i \in M, \exists v_j \in A : d_{j,i} \leq t$ **then**
5: Using the m random labelings, estimate

$$|\hat{\mathcal{M}}_i(A, M, t)| \approx \frac{m-1}{\sum_{j=1}^{m} l_j^*}$$

6: **else**
7: Set $|\hat{\mathcal{M}}_i(A, M, t)| = 0$
8: **end if**
9: **end for**
10: Compute the estimate for the exposure by averaging over all malicious neighborhood estimates

$$\hat{\chi}(A, M, t) = \frac{1}{n} \sum_{i=1}^{n} |\hat{\mathcal{M}}_i(A, M, t)|$$

11: **return** $\hat{\chi}(A, M, t)$

sequently we can determine an estimate for $|\mathcal{M}(A, M, t)|$ using the unbiased maximum likelihood estimator

$$|\mathcal{M}(A, M, t)| \approx \frac{m-1}{\sum_{j=1}^{m} l_j^*}.$$

A pseudocode implementation for the overall exposure estimation algorithm is given in Algorithm 6.

Note that in contrast to the regular influence estimation algorithm presented by Du et al. [28], we have to add an additional check in line 4: the regular least label list $l^*(j)$ alway contains v_j itself as well and will therefore return a least label for any distance $d \geq 0$ supplied as distance threshold. This is not true for the malicious least label lists l_m^*: since they only contain malicious nodes, malicious least label lists might not return any label for a given distance threshold t. This, however, directly implies that for all malicious nodes $v_j \in M$ and all source nodes $v_i \in A$ that $d_{i,j} > t$ and therefore $|\mathcal{M}(A, M, t)| = 0$. This check does not produce any computational overhead as the alternative case (line 7) automatically applies if the binary search through the malicious label list produces no result.

The time-complexity of Algorithm 6 is, given n distance network samples and m labelings of each distance network, $O(n \cdot m \cdot \log\log|M|)$. Note that, while m and n are constants in theory, they will have a notable influence on running time in practice. We will therefore list them asymptotic time complexities stated in the remainder of the paper. Choosing the right values for n and m constitutes are trade-off between running time and accuracy of the estimation.

As we will see further below, we will want to save the computed malicious least label lists for multiple exposure estimations in order to speed up our exposure minimization algorithm. Since each malicious least-label-list has an expected size of $O(\log|M|)$, the overall space requirement is $O(n \cdot m \cdot |V| \log|M|)$.

7.5.3 Exposure Estimation Accuracy

In their paper, Du et al. prove a lower bound for the sample complexity required to achieve an estimation error of ϵ with probability $1 - \delta$ for their regular exposure estimation [28, Theorem 1]. The only requirement for their proof is that the influence estimate is unbiased estimate for the mean of an exponential distribution. Since this holds true for our exposure estimate as well, and since we use the exact same sampling strategy, the same sample complexity holds for our exposure estimation algorithm as well.

Lemma 4. *Given n samples of random transmission times with*

$$n \geq \frac{C\Lambda}{\epsilon^2} \log\left(\frac{2|V|}{\delta}\right)$$

where

$$\Lambda = \max_{A:|A|\leq C} 2\frac{\chi(A,M,t)}{m-2} + 2Var(|\hat{\mathcal{M}}_i(A,M,t)|) + \frac{2a\epsilon}{3}$$

with $|\hat{\mathcal{M}}_i(A,M,t)| \leq a$, and m random labelings for each of the n transmission time samples, it holds that

$$|\hat{\chi}(A,M,t) - \chi(A,M,t)| \leq \epsilon$$

for all A with $|A| \leq C$ with probability at least $1 - \delta$.

For the network types that we consider for our evaluations (cf. Section 7.6), the relative estimation error is already below 0.01 for $n = 10000$ and $m = 5$ [28] which is why we choose these parameter values in our evaluations.

7.5.4 Exposure Minimization

The exposure minimization algorithm for the MAXIMUM-k-PRIVACY problem we developed in Chapter 6 uses the exposure estimation developed above as one of its major building blocks. We refer the reader to Section 6.7 for a detailed description of the minimization algorithm. The peusdocode implementation of the algorithm is repeated in Algorithm 7. The additive approximation error of our exposure estimation unfortunately also affects the accuracy of our exposure minimization. In the first step, it leads to an additive error when optimizing the modular approximation.

Lemma 5. *Let the supergradient $g_Y(v)$ be approximated for all $v \in V$ uniformly with error ϵ and confidence $1 - \delta$. Then the greedy mechanism finds a set $\hat{X}$ of size $|\hat{X}| = k$ with $g_Y(\hat{X}) \leq \min_{X\subseteq V:|X|=k} g_Y(X) + 2k\epsilon$ with probability at least $1 - \delta$.*

Algorithm 7 Approximating Maximum-k-Privacy

Require: Instance F, M, k of MAXIMUM-k-PRIVACY
1: $\mathcal{C} \leftarrow \{X \subseteq F \mid |X| = k\}$
2: Select random candidate solution $X^1 \in \mathcal{C}$
3: $t \leftarrow 0$
4: **repeat**
5: $\quad t \leftarrow t + 1$
6: $\quad X^{t+1} \leftarrow \arg\min_{X \in \mathcal{C}} m^{g_{X^t}}(X)$
7: **until** $X^{t+1} = X^t$
8: **return** X^t

Proof. The correctness of this statement follows from a simple pigeonhole argument made, e.g., in [82] for greedy maximization. Let $X^* = v_1, \ldots, v_k$ be the optimal solution to minimizing the supergradient $g_Y(v)$. As discussed in Section 6.7.4, it holds that

$$g_Y(v_1) \leq \ldots \leq g_Y(v_k) \leq g_Y(v)$$

for all $v \in V \setminus X^*$. Given the exposure estimation error of ϵ, the maximum error for the supergradient minimization will be observed if for all all $v_i \in X^*$, the approximated supergradient $\hat{g}_Y(v_i) = g_Y(v_i) + \epsilon$, and there exist k nodes $\hat{v}_i \in V$ such that $\hat{g}_Y(\hat{v}_i) = g_Y(\hat{v}_i) - \epsilon$ and

$$\hat{g}_Y(\hat{v}_1) \leq \ldots \leq \hat{g}_Y(\hat{v}_k) \leq \hat{g}_Y(v_i)$$

for $i \in [1, k]$. We thus get a new approximate solution $\hat{X}$ for which it holds that $g_Y(\hat{X}) \leq \min_{X \subseteq V : |X| = k} g_Y(X) + 2k\epsilon$ since for each $\hat{v}_i$, it holds that $g_Y(\hat{v}_i) \leq g_Y(v_j) + 2\epsilon$ for all $j \in [1, k]$. □

By the proof of the approximation guarantees of the majorization-minimization algorithm [53], this additive error is then inherited for the influence minimization as well. Since we can choose n to keep the additive error small, we will omit this additive error for the remainder of the paper to simplify the presentation. In our evaluation in Section 7.8 we will furthermore see that the theoretical approximation bounds do not produce any useful bounds in practice.

While Iyer et al. provide a theoretical upper bound on the maximum number of iterations that will be performed by algorithm 7, they argue that, in practice, the solution will be barely improve after a constant number of iteration (5-10) [53].

7.5.5 Computation Phases in Exposure Minimization

As we saw in Section 7.5.2, the exposure estimation can be performed in near constant time ($O(\log\log|M|)$), however only after all distance network samples, labelings and malicious least label lists have been computed. In practice, many of these more expensive computations can be performed ahead of time. We therefore partition Algorithm 6 for the exposure estimation in different phases and analyze their requirements in terms of data as well as how often they have to be computed. This partition is also listed in Figure 7.1.

Init():
Once per network:

- sample n distance network N_i from $N = (V, \alpha)$
- sample m random labelings for each N_i

Time & Space Complexity: $O(n \cdot |E| + n \cdot m \cdot |V|)$

Update():
Once per malicious node set M:

- Compute malicious least label lists for each N_i

Time Complexity:
$O\left(n \cdot m \cdot (|E| \log|M| + |V| \log|V| \log|M|)\right)$
Space Complexity: $O(n \cdot m \cdot |V| \log|M|)$

Query():
For every minimization:

- compute exposure estimates using malicious least label lists

Time Complexity: $O(n \cdot m \cdot \log \log M)$

Figure 7.1: The three phases of exposure estimation and minimization given a network $N = (V, \alpha)$ with edges $E = \{\alpha_{i,j} \in \alpha \mid \alpha_{i,j} > 0\}$ and a malicious nodes set M

Initialization Phase The initialization phase **Init()** has to be performed only once per network: it samples the n distances networks N_i that are drawn from the diffusion network $N = (V, \alpha)$, and creates the m labelings for each of the distance networks. The same distance networks and labelings can be used for exposure estimation queries with respect to any set of malicious nodes M. This computation can therefore be completely done offline and prepare the data required for the subsequent phases. Since we need to save the distance values as well as m labelings for each of the n distance networks, this pre-computation requires $O(n \cdot |E| + n \cdot m \cdot |V|)$ space, where $E = \{\alpha_{i,j} \in \alpha \mid \alpha_{i,j} > 0\}$.

Update Phase The **Update()** phase computes the malicious least label lists using the modified Algorithm 5. This has to be repeated once for every malicious node set M with respect to which we want to estimate the exposure $\chi(A, M, t)$. Depending on how the exposure estimation and minimization is used in practice, this computation can also be performed ahead of time to save on computation during run time: for instance, if a service provider offered the exposure minimization as a service, they could offer pre-defined malicious node sets such as

A) all nodes that are not friendly to the user,

B) all nodes with a minimum distance to the user,

C) all nodes that are located in a specific country,

D) all nodes that fulfill any other predicate that can be determined ahead of time.

If the malicious node set is defined dynamically during run-time, on the other hand, the malicious least label lists have to be recomputed each time a query is performed. In this case the overall query time will drastically increase.

Query Phase The query phase simply uses the malicious least label lists computed in the update phase to estimate or minimize exposure. With fixed malicious node sets for which the malicious least label lists can be computed ahead of time, we can subsequently reach the near constant $O(\log\log|M|)$ for each exposure estimation query.

In our experiments, we will evaluate the computation times for each of these phases separately. We will see that in particular the **Init()** and **Update()** phases will dominate the required computation times. Therefore, having pre-defined malicious node sets might drastically improve the practicality of this approach.

7.5.6 Greedy Heuristics for Exposure Minimization

In addition to the XpoMin approximation algorithm discussed above, we also use a greedy heuristic to minimize the exposure of an initial infection: we assign each node in the friendly node set F an objective value, sort them by their objective value and return the top k nodes. This greedy heuristic is parametric in the objective values that is used to sort the nodes. In our framework we consider the following objective functions:

1) **Marginal** sorted by marginal contribution to the exposure in ascending order.

2) **Singular** sorted by singular contribution to the exposure in ascending order.

3) **Shortest Path** sorted by shortest path to any of the malicious nodes in descending order using unit edge distances for all edges.

4) **Degree** sorted by node degree in ascending order

5) **Malicious Degree** sorted by number of edges that do not lead into either user or friendly nodes in ascending order.

Both the marginal as well as singular heuristic rely on the exposure estimation algorithm we developed above and inherit the corresponding time complexities. For the shortest path heuristic, we need to run the a single source shortest path for all friendly nodes, which leads to a $O(|F|(|E| + |V|\log|V|))$ time complexity using Dijkstra's algorithm with Fibonacci heaps.

In the case of the degree and malicious degree heuristics we need to touch each edge incident to friendly nodes once. In the worst case this corresponds to a $O(E)$ worst case complexity. Since all objectives are real or integer valued, the sorting can be performed

in linear time as well. In practice, the computations required for the shortest path, degree and malicious degree heuristics can be performed offline, leading to a simple constant time lookup during runtime.

7.5.7 Influence vs. Exposure

Some of the heuristics above ignore any information about the malicious nodes provided by the MAXIMUM-k-PRIVACY instance, and in our evaluations we will see that in particular the degree heuristics perform very well for a set of randomly chosen malicious nodes. Since the set of malicious nodes separates the notion exposure from the notion of influence, one could argue that minimizing influence alone can be enough to also minimize exposure. In the following we provide some observations with respect to the relation between exposure and influence.

First, it is very easy to to construct MAXIMUM-k-PRIVACY instances where the degree heuristics will achieve an arbitrarily bad approximation result. Consider, for instance, two friendly nodes v_i and v_j, where the node degree d_i of v_i is much larger than the node degree d_j of v_j. Further, let G_i and G_j be the disconnected subgraphs that connect only to v_i and v_j respectively. Assuming we want to find the node with minimum exposure, the degree heuristics would always choose v_j as the solution. However, if all of G_j is malicious, while G_i is not, we can easily scale G_j so that the approximation is arbitrarily bad.

Lemma 6. *For any $r \in [0, \infty)$, there is a* MAXIMUM-k-PRIVACY *instance with optimal solution F^* and approximate solution $\hat{F}$ computed by the degree heuristic such that*

$$\frac{\chi(\hat{F}, M, t)}{\chi(F^*, M, t)} > r.$$

If the malicious nodes are chosen uniformly at random, however, minimizing influence among the non friendly nodes can be enough to also minimize exposure in expectation.

Lemma 7. *Given an instance of the* MAXIMUM-k-PRIVACY*, let the malicious node set M be given by choosing any node $v \in V$ with probability p. Then*

$$\operatorname*{arg\,min}_{F' \subseteq F, |F'|=k} \sigma(F'; t) = \operatorname*{arg\,min}_{F' \subseteq F, |F'|=k} \mathbb{E}[\chi(F', M, t)].$$

Proof. By definition,

$$\sigma(F; t) = \sum_{v_i \in V} \Pr[t_i \leq t \mid F]$$

and

$$\chi_N(F, M, t) = \sum_{m_i \in M} \Pr[t_i \leq t \mid F] = \sum_{v_i \in V} X_i \cdot \Pr[t_i \leq t \mid F],$$

where X_i is a binary random variable that indicates whether v_i is in the malicious node

set. Since the malicious nodes are chosen uniformly with probability p, it holds that

$$\begin{aligned}\mathbb{E}[\chi(F', M, t)] &= \sum_{v_i \in V} p \cdot \Pr[t_i \leq t \mid F] \\ &= p \sum_{v_i \in V} \Pr[t_i \leq t \mid F] \\ &= p \cdot \sigma(F; t).\end{aligned}$$

□

Therefore, if no particular malicious node set can be determined for the MAXIMUM-k-PRIVACY, minimizing influence instead of exposure might be the better choice, in particular because this saves the **Update()** computation phase that depends on the considered malicious node set. As a consequence we will in particular observe that the node degree heuristics work very well in our evaluations since we randomly choose the malicious node sets. In practice, however, the set of malicious nodes would be chosen by specific node attributes, e.g. country of origin. In these cases, a strict exposure minimization with the full exposure minimization algorithm will show better results.

Unfortunately, to our knowledge there currently does not exist a suitable data set that would allow for a comparative evaluation our of approximation algorithms under non-random malicious node sets. The collection of a suitable data set and subsequent evaluation of our algorithms would therefore be an important step for future work.

7.6 Experimental Setup

In this section, we provide details for our implementation of XpoMin, as well as introduce the general setup for our evaluations.

7.6.1 Implementation & Hardware

We implemented XpoMin in C++ using the C++11 standard. We rely on the SNAP library for basic graph operations such as shortest paths and degree computations [69]. For the exposure estimation itself, we modify the implementation of the influence estimation algorithm [28] in PtPack, a C++ multivariate temporal point process library[2]. The implementation of XpoMin along with the data sets used for our evaluations will be made available publicly after the publication of paper corresponding to this chapter [P4]. All evaluations were performed on Dell PowerEdge R820 servers with 64 virtual cores at 2.60GHz each and 768GB of RAM.

7.6.2 Data Sets

We evaluate the performance of XpoMin using synthetic as well as real diffusion network data. The synthetic networks allow us to cover a larger array of network configurations since there is barely any real networking data on information diffusion available in the literature.

[2] https://github.com/dunan/MultiVariatePointProcess

Synthethic Data For our synthetic networks, we use Kronecker graphs which have successfully been used in the literature for the modeling and analysis of social and other comparable networks. Leskovec et al. [67] in particular prove that Kronecker graphs produce the same network properties that we typically observe from real networks. Kronecker graphs are a parametric model that use a 2x2 initiator matrix to define various types of networks. In our evaluations we consider three different network types:

A) core-periphery networks (parameter matrix [0.9 0.5; 0.5 0.3]), which model information diffusion in real networks [68],

B) hierarchical networks (parameter matrix [0.9 0.1; 0.1 0.9]]), which model web graphs and biological networks [19],

C) and random networks (parameter matrix [0.5 0.5; 0.5 0.5]), which model network used in physics and graph theory [34].

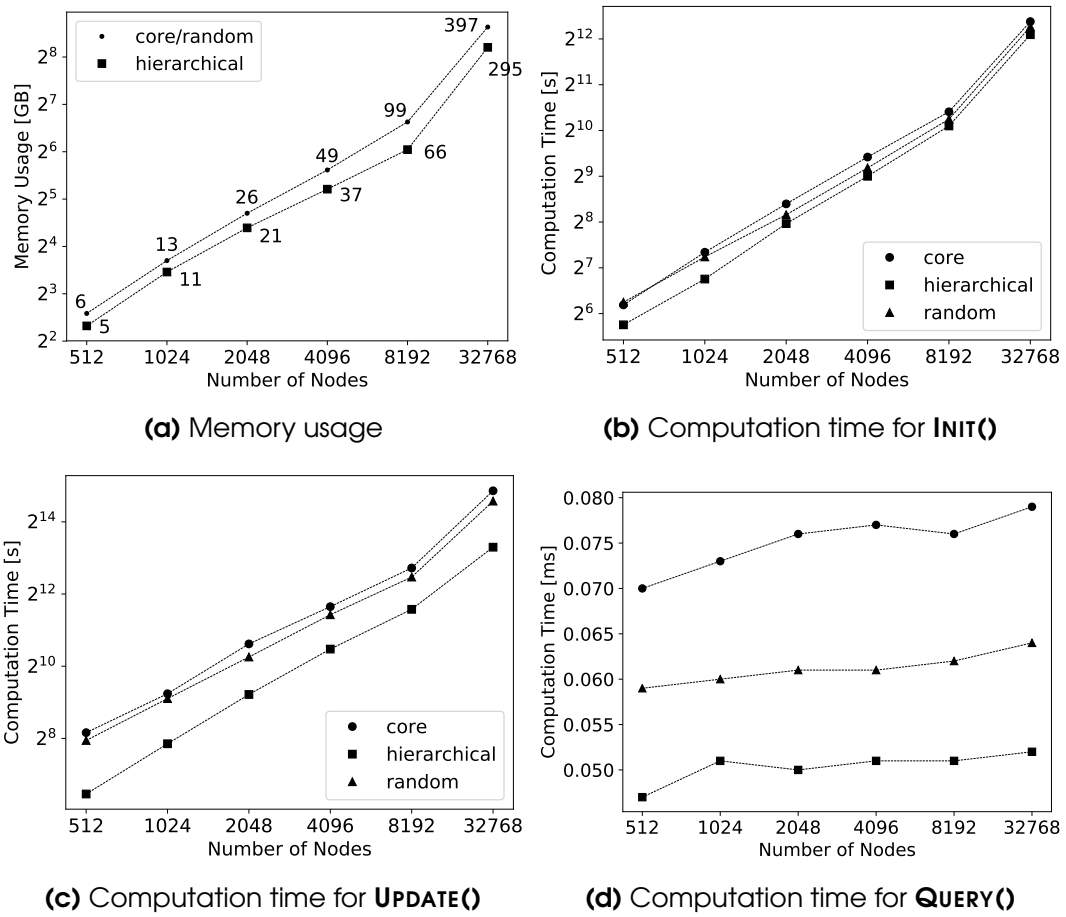
We then assign transmission parameters to each edge in the resulting networks. We sample each transmission parameter uniformly at random in the range from 0.5 to 1.5, a process which has successfully been used in the past for the analysis of influence maximization algorithms [46]. To generate these graphs, we utilize the implementation of the graph generation algorithm made available publicly by Gomez-Rodriguez[3]. Due to its implementation, all synthetic graphs have a node number that is a power of two.

Real Data The Memetracker data set is a collection of web pages published by one million online domains in the time from October 2008 to April 2009 [66]. The data set also contains hyperlinks mentioned on each of the webpages that point to other webpages in the data set. It has therefore thoroughly been used in the literature for the analysis of diffusion processes and other networking properties in the web.

In their paper in which they introduce this data set, Leskovec et al. also develop the phrase clustering methodology which reduces the content of each webpage to central phrases. Similar phrases are subsequently combined into phrase clusters to capture that the same information is expressed through slightly different phrases. The resulting phrase-cluster data set provides a list of webpages in which a phrase of a phrase-cluster appears in. This list is sorted by the time of publication of each web page and thus allows us to follow the flow of information between webpages in time.

For our evaluations, we extract the top 1000 domains with the most published webpages and build information diffusion cascades between these by following the phrase-cluster trail observed in the phrase-cluster data set. We then use the 10000 longest cascades to infer the underlying diffusion network using NetRate [44]. Since the average transmission times in this network are significantly larger than what we sampled for our synthetic networks, we normalize the inferred transmission parameters into a similar range as used in the synthetic case. Finally, we remove edges with very small transmission rates ($<< 10^{-2}$) since such edges have a negligible likelihood of creating a contribution to the exposure within the time thresholds that we consider in our experiments.

[3] https://people.mpi-sws.org/ manuelgr/influmax/



(a) Memory usage

(b) Computation time for **INIT()**

(c) Computation time for **UPDATE()**

(d) Computation time for **QUERY()**

Figure 7.2: Performance of XpoMin for **INIT()**, **UPDATE()** and **QUERY()** phases as well as memory usage across multiple networks.

7.6.3 Pre-Processing

To ensure a non-zero curvature κ for all of the MAXIMUM-k-PRIVACY instances that we consider, we perform a pre-processing of each instance in which we remove all friendly nodes that do not have a path to any malicious node. Any such node $v_j \in F$ has a singular contribution $\chi(v_j, M, t) = 0$ for any time threshold t and will therefore never cause the diffusion of information to a malicious node.

7.6.4 Evaluation Methodology

All of the results reported in the following sections are averaged over multiple runs of the exposure minimization algorithm on the synthetic and real diffusion networks described above. To keep our evaluations tractable, we scale the number of instances we compute with the size of the networks. For networks up to $2^{11} = 2048$ nodes, we will compute 2000 different instances, while for networks larger than 2^{11} and with a size up to $2^{13} = 8192$ we will compute 200 different instances. Finally, we compute 40

instances on synthetic graphs with $2^{15} = 32768$ nodes to obtain a rough estimate of XpoMin's performance for larger graphs.

To generate these instances, we consider several instance parameters: first, we iterate over a number of different user nodes v_i which we will consider as the source of the information that is shared. We always consider the nodes adjacent to this source as the set of friendly nodes F.

Next, we vary the time threshold t for which we minimize the exposure $\chi(A, M, t)$. Since we sampled the network transmission parameters for our synthetic networks uniformly from the range $[0.5, 1.5]$, we chose the time thresholds $t \in \{0.5, 1.0, 2.0, 10.0\}$ to emulate short, intermediate and long time horizons. We also vary the utility threshold k, which defines the number of friendly nodes that we want to share information with, in the range $k \in \{1, 2, 5, 10\}$. Since, on average, the number of F in the networks we consider will be smaller than 10, these parameters cover a wide range of the potentially possible queries.

Finally, we use various rules for generating the malicious node sets:

1) all nodes but F and the user v_i

2) all nodes with minimum distance 2 from v_i and F

3) randomly chosen from $V \setminus (F \cup \{v_i\})$ with probability $p \in \{0.1, 0.3, 0.5\}$

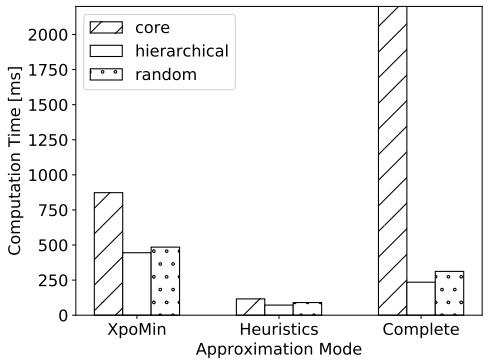
This instance generation methodology could potentially allow to perform a deeper analysis in how time thresholds and different malicious node sets impact the minimization performance. In this work, we will average our results over all instances to keep the exposition simple. A deeper analysis could, however, prove a interesting direction for future work.

7.7 XpoMin Performance Evaluation

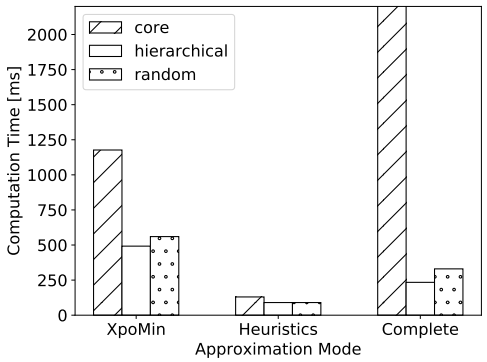
To evaluate the performance of XpoMin, we separately measure the time required for the **Init()** and **Update()** phases while running the exposure minimization. Since the time required for the **Query()** phase typically is very small, we infer it from dividing the time required to estimate the singular contributions of every friendly nodes in a problem instance by the number of friendly nodes. For the memory, we record the peak memory usage for minimizing each of the instances and average the observed values. Since we are only interested in the evolution XpoMin's performance with network size, we only consider synthetic networks for this evaluation.

We present our results by network size as well as network type (core-periphery, hierarchical, random) since the different network types result in a different amount of edges: while the number of edges is nearly the same for both core-periphery and random networks, with an average node degree of 2, hierarchical Kronecker networks produce around 25% less edges. As we will see below, this will have a notable impact on the performance of XpoMin, as was also predicted by the theoretical complexity analysis (cf. Chapter 6).

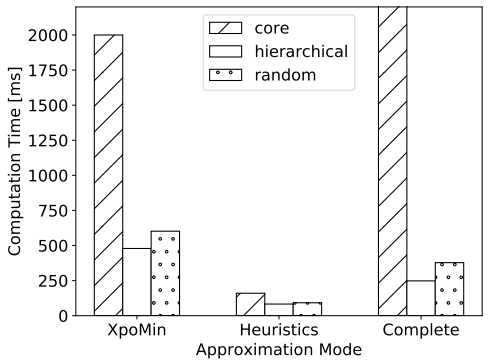
The results of our performance evaluations are shown in Figure 7.2. Since we double the number of nodes for each increase in network size, we present the results with a



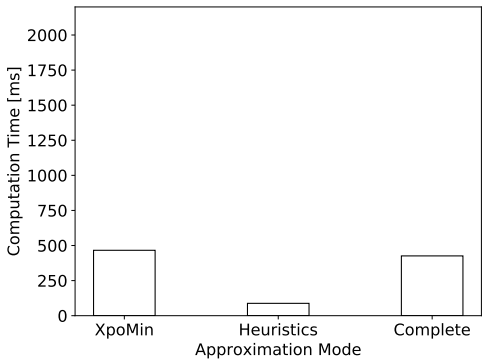
(a) Minimization time on 512 node networks.



(b) Minimization time on 1024 node networks.



(c) Minimization time on 2048 node networks.



(d) Minimization time on phrase cluster networks.

Figure 7.3: Computation time for exposure minimization across multiple networks.

logarithmic scale on the y-axis. Consequently, a linear slope of in these diagrams also represents a linear increase in reality. Please keep in mind that, since we only considered networks of size $3^15 = 32768$ to represent larger graphs, we essentially skip a data point at the higher end of the diagrams. This will typically results in a doubling of the slope that we observe between the penultimate and ultimate data point.

For the memory usage, Figure 7.2a shows its linear dependence on the number of nodes in the network. Since the memory usage of both, core-periphery and random network was nearly identical in our experiments, we consolidated them in this illustration. For hierarchical networks, however, we can see a notable decrease in the used memory, explained by the smaller number of edges that are present in such networks. Note that all of these measurements were done with $n = 10000$ distance network samples and $m = 5$ labelings. The required memory would linearly increase in each of these parameters should you choose to increase them (for improved accuracy of the exposure estimation). In Figures 7.2b and 7.2c, we show the time required for the **Init()** and **Update()** phases in seconds. As in the case of memory usage, we can see a linear increase in runtime for both of these phases with increasing network size. Again, the lower number of edges for hierarchical networks results in a smaller computation time

for such networks across the board. On average, the **Update()** time is around four times as large as the **Init()** time: in **Init()**, we only sample network edge weights and node labels, whereas for **Update()** we also have to compute the malicious least label lists.

As we can see, trying to recompute the malicious least label lists during runtime to allow for dynamically generated malicious node sets is not very practical. An adaptation of the XpoMin framework in practice would instead massively benefit from using predefined malicious node sets to allow for a offline pre-computation of the malicious least label lists.

Figure 7.2d shows that, In contrast to the **Init()** and **Update()** times, the **Query()** times remain mostly constant across network sizes. While we can again see a smaller computation time for hierarchical graphs, this time this is not due to the lower amount of edges: as our theoretical analysis in Section 7.5.2 showed, the running time of a exposure estimation query depends on the length of the malicious least label lists, which in expectation is logarithmic in the malicious node set size. However, the specific structure of each network can cause the size of the malicious least label list to increase or decrease. We leave the analysis of the impact of the network type on the length of the malicious least label lists as an interesting direction for future work.

In absolute values, the **Query()** times are in the range of $5 - 7 \cdot 10^{-2}$ms, which allows for a near instant response time in practice. Since the **Query()** computation time linearly depends on n and m, even increasing these constant by a order of magnitude will only result in computation times in the millisecond range which can still be sufficient in practice, and allow for much higher accuracy of the exposure estimation.

7.7.1 Parallelization

The results show that for a network with $2^{15} = 32768$ nodes, we are already taking around 4.5 hours to compute the malicious least label lists and and around 1 hour for the **Init()** phase. Scaling this up to a million node network, for instance, this results in a running time of around 140 hours or 6 days for the **Update()** and 32 hours for the **Init()** phases. However, this is purely single-thread performance, and as discussed at the end of Section 7.5, we can straightforwardly apply parallelization to improve upon these values.

While we were not able to perform extensive evaluations with parallelization due to time and memory constraints, first runs on the $2^{15} = 32768$ node network with 32 threads show promising results. The required overall computation times decrease nearly linearly in the number of available threads, going down to around 10 minutes for the **Update()** phase and 3 minutes for the **Init()** phase. Interpolating to a million node network, this yields a running time of around 6.3 hours for **Update()** and 2.5 hours for **Init()**.

7.7.2 Minimization Performance

Ultimately, the main purpose of XpoMin is to minimize the exposure for instances of the Maximum-k-Privacy problem. we therefore also evaluated the computation time of the approximation algorithm we presented in Section 7.5. To simplify notation, we

will, in the following, denote this approximation algorithm with the name XpoMin as well.

We compare the computation time of XpoMin with the other heuristics we presented in Section 7.5. Since each of these heuristics simply sorts the friendly nodes F by their respective objective values which have been pre-computed, their running times were the same across the board in our evaluations. We therefore combine them into one entry in our diagrams. We also add the computation time required for the complete search that iterates over all possible friendly subsets $F' \subseteq F$ with $|F'| = k$.

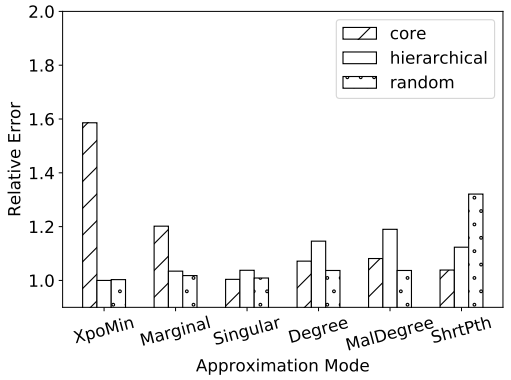
The evaluation results for the exposure minimization divided by network size and network type are presented in Figure 7.3. Across all network sizes, the exposure minimization using heuristics is clearly the fastest at around 100ms. Notably, this running time stays relatively constant across network sizes, indicating that the average node-degree stays the same. We do, however, see an increase for core-periphery networks compared to the other network types which can be explained by the increased time for **Query()** that we observed above.

For the XpoMin approximation algorithm, we can observe a computation time that is generally at least four times longer compared to the simpler heuristics. In particular the core-periphery networks seem to cause significant issues to the approximation algorithm since they cause its running time to increase linearly with the network size, up to nearly 2s for the 2048 node core-periphery network. As we will see in Section 7.8, this is also reflected in XpoMin's minimization accuracy on core-periphery networks where it performs especially poorly.

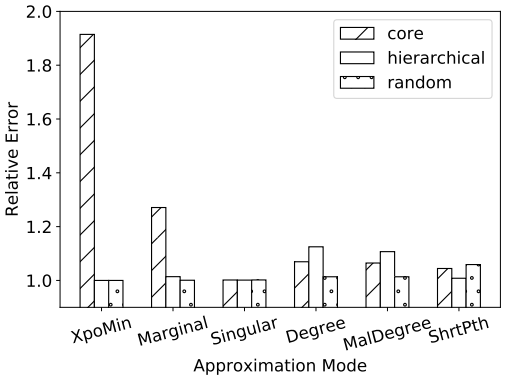
The observed computation times for the complete search show two tendencies: for hierarchical and random Kronecker networks, as well as the real phrase cluster network, it shows a computation time between the time required for the heuristics and the XpoMin approximation algorithm. This is explained by the typically small number of friendly nodes encountered in such network. For core-periphery networks, however, the average computation time increases dramatically, which is caused by a few users with a very large number of friendly nodes. In such cases, the complete search takes several hours (to even days) to complete. Consequently, the complete search can be a meaningful approach to to exactly minimize the exposure given that we can detect instances with a large number of friendly nodes and use the approximate approaches instead.

7.8 Exposure Minimization Accuracy

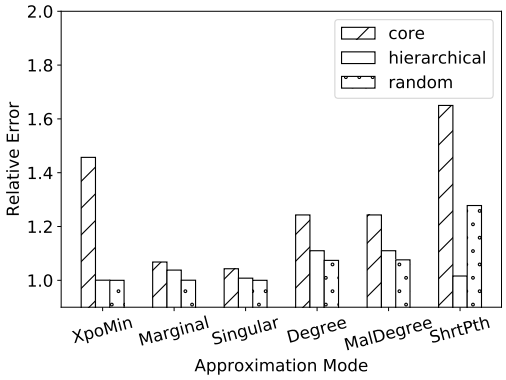
To evaluate the accuracy of our influence minimization algorithms, we compare their output with the output of a complete search that we described above. Our evaluation metric will be the *relative error* achieved by our approximation algorithms. That is, given the approximate minimum exposure $\hat{\chi}^*$ computed by the approximation algorithm and the actual minimum exposure $\chi*$ computed by the complete search, we compare the relative error $\delta = \frac{\hat{\chi}^*}{\chi*}$ of each approximation algorithm. Naturally, the relative error is always greater than one, and the smaller it is the better the approximation algorithm approximates the minimum exposure.



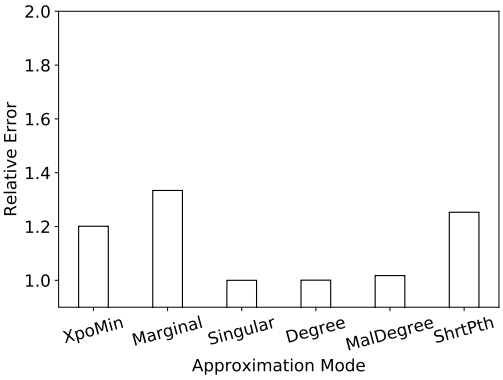
(a) Synthetic networks with 512 nodes.



(b) Synthetic networks with 1024 nodes.



(c) Synthetic networks with 2048 nodes.



(d) Phrase cluster network.

Figure 7.4: Average relative error of XpoMin (cf. Section 7.5.4) and other minimization heuristics (cf. Section 7.5.6).

Network Type	Average Curvature
core-periphery	$8.73 \cdot 10^{-4}$
hierarchical	$8.28 \cdot 10^{-2}$
random	$8.34 \cdot 10^{-3}$
phrase cluster	$1.85 \cdot 10^{-4}$

Table 7.1: Average curvatures κ for the four considered network types.

7.8.1 Curvature in Practice

Before we begin with the empirical evaluation of the relative error on our synthetic and real diffusion networks, we first take a look at the average curvature κ (cf. Chapter 6) of our problem instances. Recall that always $\kappa < 1$ and, given that $\kappa > 0$, the XpoMin approximation algorithm provides a $\frac{1}{\kappa}$ approximation bound for the minimization. In our evaluations, these curvature values are extremely small across the board, and consequently only allow for a very bad theoretical approximation bound for the XpoMin algorithm. Fortunately, our empirical evaluations show that, in practice, XpoMin and the other minimization heuristics we discussed in Section 7.5 achieve much better re-

sults. A listing containing the average curvature values for each network type can be found in Table 7.1.

7.8.2 Relative Error of Approximation

These evaluation results for the relative error are displayed in Figure 7.4. Overall we obtain a maximum relative error < 2 which shows that the all approximation algorithms work much better than indicated by the theoretical worst case bound we derived above. While the XpoMin algorithm performs best for the hierarchical and random Kronecker networks, with nearly always reaching the optimal solution, its shows comparatively poor performance for the core-periphery networks, reaching the maximum observed relative error for the core-periphery network with 1024 nodes. Coincidentally, we can also observe a comparatively poor performance by the marginal heuristic for the same cases: since XpoMin uses a combination of overall and intermediate marginal contributions for its optimization (cf. Section 7.5.4), this might indicate that, in particular for core-periphery networks, spread between singular and marginal contributions is especially large, leading to a bad results.

Looking at the relative errors for each instance separately, we were able to see that the high average relative error for XpoMin on core-periphery networks is caused by a small number of instances with a huge relative error around 100 that dominate the average. Cutting off the worst 20% of the instances, we obtain an average relative error much closer to the remaining cases.

The heuristic approaches, in particular the singular heuristic, perform very well across all problem instances in synthetic and real networks. It therefore seems that, in realistic diffusion networks, the exposure function very much behaves like a linear function instead of submodular, allowing us to minimize the exposure by just choosing the nodes that have the smallest singular exposure contribution.

Among the transmission rate agnostic heuristics, the shortest path heuristic shows comparatively bad approximation errors in multiple instances. Combined with the increased pre-processing cost of computing the shortest paths between all friendly and malicious nodes, this makes the shortest path heuristic rather undesirable in practice.

Surprisingly, both the degree and malicious degree (cf. Section 7.5.6) heuristics perform comparatively well while using the minimum amount of information for the optimization. In particular for the real phrase cluster network, the degree heuristics almost always achieve the optimal solution. The only other heuristic that achieves the same is the singular heuristic, which however requires access to the exposure minimization algorithm to perform the minimization.

Since these transmission rate agnostic measures do not consider the malicious node set in their minimization, these results show that a simple influence estimation is sufficient to also minimize exposure given the malicious node set is chosen randomly, as we have shown in Section 7.5.7. It is, however, to be expected that under non-random malicious node sets, these network agnostic measures would not perform as well.

7.9 Discussion

Our evaluations show that the XpoMin framework allows for accurate exposure minimization in practice despite very bad theoretical approximation bounds given by the submodularity of our objective functions. As already dicussed in Section 7.5.1, however, these algorithms are not yet suited for user-side application on mobile devices or PCs with limited computational power, caused by the required information about the whole considered network. Instead, this implementation can be seen as a first step towards exposure minimization in practice, providing a working approach for deployment on the side of the service provider, or to be used as a benchmark for future developments.

The application of exposure minimization in practice would typically not consider singular malicious nodes that fulfill very specific properties: in the presence of an adversary that is actively looking to learn information about the user, the expected exposure estimation produced by our framework will not provide any meaningful guarantees. Instead, we see the application of XpoMin in empowering the user to control how far their shared information is likely to propagate through the network.

Similar approaches have been discussed in the past: Backes et al. [7], for instance, propose a mechanism to enforce an expiration date on media shared on the Internet. This mechanism and other related mechanisms that try to control data lineage and to provide some means of access control to shared information, however, rely heavily on some trusted third-party to mediate the access to the information. This trusted third party assumption make these approaches very unlikely to be adopted in practice [78]: its adds additional overhead to any action performed by users and adds an additional point of failure.

The exposure minimization approach with XpoMin is completely free of this trusted third party assumption. Consequently, however, XpoMin also does not provide any hard guarantees about the propagation of information, but instead relies on statistical arguments only. It can therefore not be seen as a privacy mechanism that enforces privacy, but instead as an advisory tool that can empower the user to make privacy-conscious decisions.

7.10 Conclusion

In this chapter, we present XpoMin, a framework for the efficient exposure estimation and minimization in continuous-time diffusion networks. To achieve an efficient and practical exposure estimation algorithm, we adopt the sampling based influence estimation algorithm put forward by Du et al. [28]. We show that, with slight modifications to their algorithm, we can achieve an efficient estimation of exposure, which is a generalization of influence. We also present a natural partitioning of this algorithm into three phases: the **Init()**, **Update()** and **Query()** phases. Depending on the use case, the majority of the required computation for the exposure estimation can be performed offline within the **Init()** and **Update()** phases, allowing for near constant query times during run time. Our performance evaluation of XpoMin demonstrate the linear dependence of the **Init()** and **Update()** phases on the size of the network, and show that the **Query()** phase can provide a near constant time exposure estimation across

the full range of network sizes that we consider.

In a second step we evaluated the accuracy of XpoMin's exposure minimization on realistic networks. To this end we implement the approximation algorithm we presented in Chapter 6 as well as several greedy heuristics. Our evaluations show that, despite very bad theoretical worst case guarantees, we can achieve an average relative error of $\delta < 2$ in the worst case for the approximation algorithm. Furthermore, the singular and degree greedy heuristics work very well, achieving the actual optimal result in the majority of the cases.

In particular the diffusion model agnostic greedy heuristic give hope for the adaptation of XpoMin by individual users who often do not possess knowledge about the whole network they are interacting with. Overall, we pave the way for a practical approach to exposure minimization, providing users the tools necessary to make informed decisions about their information sharing behavior.

7.11 Future Work

The implementation presented here only constitutes a first step towards solving the problem of exposure minimization in social networks and several directions for future work are apparent. First, in order to make the exposure minimization viable completely on user side, one has to evaluate how exposure estimation and minimization with a limited view of the network compares to the baseline solution provided in this work: in practice, a single user will not be able to access or know the while network he interacts with, and instead will have to make his decisions based on limited information given by their local view of the network.

A second direction involves the type of information that is shared. As we discussed in the related work in Chapter 3, there already exists some approaches to modeling heterogeneous diffusion models that take into account the type of information that is shared to determine how quickly it spreads throughout the network. It is easy to imagine that certain users in the network will more likely adopt information of a certain type, while another user will more likely adopt and share information of another. Taking into the account the content, and therefore type, of information that is shared will allow for an exposure minimization better tailored to the specific use case.

Building on this second direction, it would also be interesting to see whether the exposure estimation and minimization could be performed dynamically without incurring too much computational overhead: our current solution assumes a static network that never changes to estimate the likelihood that shared information reaches certain nodes in the network. In practice, however, social networks are highly dynamic, and even the information sharing behavior might be dynamically changing, caused by changes in trends of communication and in interests of users throughout the network. If we were able to capture the dynamic information sharing behavior in our exposure estimation model, this could allow for a much more fine-grained exposure minimization.

8
Conclusion

In this dissertation we presented the results of two lines of work on privacy and anonymity in decentralized, open settings such as modern social media and social networking platforms. As we discussed in chapter 2, existing traditional privacy mechanisms are not applicable to such settings. Due to the ever increasing number of users that participate on such online platforms, a comprehensive approach to dealing with such issues is exceedingly required. To this end, new formal approaches to reasoning about privacy in decentralized, open settings are particularly necessary for well-founded reasoning on this topic and the development of suitable practical solutions in the next step.

Our main focus in this dissertation was to develop such formal foundations for reasoning about the different facets of privacy in decentralized, open settings. In particular, we wanted to develop general purpose frameworks that can be instantiated for various use cases (e.g. our anonymity framework that can be instantiated with different user models depending on which (types of) information is considered). Furthermore, we wanted to investigate whether and how insights from traditional privacy mechanisms transfer to the open settings. Another major motivation for our work was the central insight that decentralized, open settings generally do not allow for provable privacy guarantees. Instead we have to rely on sound privacy risk assessments to provide meaningful assistance to the end user in situations affecting their privacy in online social media and social networking platforms.

The results presented herein have been published across multiple publications [P2, P1, P3, P4]. Here, the first two publications [P2, P1] contributed to the development of a formal framework for information disclosure and the notion of (k, d)-anonymity. In [P2] we develop both of these and provided an initial experimental validation of the framework and the (k, d)-anonymity notion (cf. Chapter 4). To further investigate the relation between anonymity and linkability in decentralized, open settings, we then performed extensive experimental evaluations on the Reddit social media platform in [P1] (cf. Chapter 5). Our results here showed that, in contrast to the traditional closed setting, having a degree of anonymity does not necessarily imply unlinkability in decentralized, open settings. This is in line with our formal framework for information disclosure developed in [P2] that predicts the potential existence of identifying attributes outside the perfect anonymity setting.

The other two publications contributed to the development of the exposure minimization approach to controlling information diffusion in continuous-time diffusion networks [P3, P4]. In [P3] we introduce the exposure minimization approach and develop the corresponding maximization problems (cf. Chapter 6). While our analyses show that these optimization problems are NP-hard to solve exactly, we find efficient approximate solutions by leveraging the submodularity of the corresponding objective functions. A main drawback of the proposed approximation algorithm is, however, that it assumes an efficient algorithm for the exposure estimation algorithm. Since the influence estimation problem had previously been shown to be $\#P$-hard, the more general exposure minimization problem was also at least $\#P$-hard and would not easily allow for an efficient, exact solution. Prompted by this issue, we then investigated an efficient approximation of exposure estimation and the implementation of the approximate exposure minimization algorithm in [P4] (cf. Chapter 7). The presented solution

adapts a scalable approximation algorithm for influence estimation to the exposure estimation problem and then uses this new algorithm as a building block for the implementation of the exposure minimization. Due to the advantageous structure of the resulting algorithm we achieved a near constant query time for exposure minimization, given extensive pre-computations to build the necessary data structures.

Overall, we presented two novel approaches to reasoning about privacy in decentralized, open settings. Both approaches follow the idea of providing sound privacy risk assessments to assist user to make informed decisions about the information dissemination behavior. Consequently, they present a departure from traditional, provable solutions which seem very difficult to be achieved in decentralized, open settings.

Bibliography

Author's Papers for this Thesis

[P1] Backes, M., Berrang, P., Goga, O., Gummadi, K. P., and Manoharan, P. On Profile Linkability Despite Anonymity in Social Media Systems. In: *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. ACM. 2016, 25–35.

[P2] Backes, M., Berrang, P., and Manoharan, P. From Zoos to Safaris—From Closed-World Enforcement to Open-World Assessment of Privacy. In: *Foundations of Security Analysis and Design VIII*. Springer, 2015, 87–138.

[P3] Backes, M., Gomez-Rodriguez, M., Manoharan, P., and Surma, B. Reconciling Privacy and Utility in Continuous-time Diffusion Networks. In: *Proceedings of the 30th Computer Security Foundations Symposium (CSF)*. IEEE. 2017, 292–304.

[P4] Manoharan, P. and Backes, M. XpoMin: Towards Practical Exposure Minimization in Continuous Time Diffusion Networks. In: *under submission*. 2018.

Other Papers of the Author

[S1] Backes, M., Berrang, P., Humbert, M., and Manoharan, P. Membership privacy in MicroRNA-based studies. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM. 2016, 319–330.

[S2] Backes, M., Kate, A., Manoharan, P., Meiser, S., and Mohammadi, E. AnoA: A framework for analyzing anonymous communication protocols. In: *Proceedings of the 26th Computer Security Foundations Symposium (CSF)*. IEEE. 2013, 163–178.

[S3] Backes, M., Manoharan, P., and Mohammadi, E. TUC: Time-sensitive and Modular Analysis of Anonymous Communication. In: *Proceedings of the 27th Computer Security Foundations Symposium (CSF)*. IEEE. 2014, 383–397.

[S4] Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280* (2017).

[S5] Grosse, K., Papernot, N., Manoharan, P., Backes, M., and McDaniel, P. Adversarial Examples for Malware Detection. In: *Proceedings of the 22nd European Symposium on Research in Computer Security (ESORICS)*. Springer. 2017, 62–79.

Other references

[1] Abbasi, A. and Chen, H. Writeprints: A Stylometric Approach to Identity-level Identification and Similarity Detection in Cyberspace. *ACM Transactions on Information Systems* 26, 2 (2008), 7:1–7:29.

[2] Acquisti, A., Gross, R., and Stutzman, F. Face Recognition and Privacy in the Age of Augmented Reality. *Journal of Privacy and Confidentiality* 6, 2 (2014), 1.

[3] Adar, E. and Adamic, L. A. Tracking Information Epidemics in Blogspace. In: *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*. 2005, 207–214.

[4] Afroz, S., Brennan, M., and Greenstadt, R. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In: *Proceedings of the 33rd IEEE Symposium on Security and Privacy (S&P)*. 2012, 461–475.

[5] Backes, M., Berrang, P., Goga, O., Gummadi, K., and Manoharan, P. *On Profile Linkability Despite Anonymity in Social Media Systems - Supplementary Material*. `https://infsec.cs.uni-saarland.de/projects/reddit_anonymity/`.

[6] Backes, M., Berrang, P., and Manoharan, P. *From Closed-world Enforcement to Open-world Assessment of Privacy*. http://arxiv.org/abs/1502.03346. eprint arXiv:1502.03346 – cs.CR. 2016.

[7] Backes, M., Gerling, S., Lorenz, S., and Lukas, S. X-pire 2.0: A User-controlled Expiration Date and Copy Protection Mechanism. In: *Proceedings of the 29th annual ACM Symposium on Applied Computing*. ACM. 2014, 1633–1640.

[8] Backes, M., Kate, A., Manoharan, P., Meiser, S., and Mohammadi, E. AnoA: A Framework for Analyzing Anonymous Communication Protocols. In: *Proceedings of the 26th IEEE Computer Security Foundations Symposium (CSF)*. 2013, 163–178.

[9] Balduzzi, M., Platzer, C., Holz, T., Kirda, E., Balzarotti, D., and Kruegel, C. Abusing Social Networks for Automated User Profiling. In: *Proceedings of the 13th international Conference on Recent Advances in Intrusion Detection (RAID)*. 2010, 422–441.

[10] Barbieri, N., Bonchi, F., and Manco, G. Topic-Aware Social Influence Propagation Models. In: *Proceedings of the 2012 IEEE 12th International Conference on Data Mining (ICDM)*. 2012, 81–90.

[11] Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. Discriminating Gender on Twitter. In: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2011, 1301–1309.

[12] Calì, A., Calvanese, D., Colucci, S., Di Noia, T., and Donini, F. A Logic-Based Approach for Matching User Profiles. In: *Knowledge-Based Intelligent Information and Engineering Systems*. 2004, 187–195.

[13] Cecaj, A., Mamei, M., and Zambonelli, F. Re-identification and Information Fusion between Anonymized CDR and Social Network Data. *Journal of Ambient Intelligence and Humanized Computing* (2015), 1–14.

[14] Chatzikokolakis, K., Andrés, M., Bordenabe, N., and Palamidessi, C. Broadening the Scope of Differential Privacy Using Metrics. In: *Proceedings of the 13th Privacy Enhancing Technologies Symposium (PETS)*. 2013, 82–102.

[15] Chen, R., Fung, B. C., Yu, P. S., and Desai, B. C. Correlated Network Data Publication via Differential Privacy. *The VLDB Journal* 23, 4 (2014), 653–676.

[16] Chen, T., Kaafar, M. A., Friedman, A., and Boreli, R. Is More Always Merrier?: A Deep Dive into Online Social Footprints. In: *Proceedings of the 2012 ACM Workshop on Online Social Networks (WOSN)*. 2012, 67–72.

[17] Chen, W., Wang, C., and Wang, Y. Scalable Influence Maximization for Prevalent Viral Marketing in Large-scale Social Networks. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2010, 1029–1038.

[18] Cheng, Z., Caverlee, J., and Lee, K. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In: *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM)*. 2010, 759–768.

[19] Clauset, A., Moore, C., and Newman, M. Hierarchical Structure and the Prediction of Missing Links in Networks. *Nature* 452, 7191 (2008), 98–101.

[20] Cohen, E. Size-Estimation Framework with Applications to Transitive Closure and Reachability. *Journal of Computer and System Sciences* 55, 3 (1997), 441–453.

[21] Conover, M. D., Goncalves, B., Ratkiewicz, J., Flammini, A., and Menczer, F. Predicting the Political Alignment of Twitter Users. In: *Proceedings of the 3rd IEEE Inernational Conference on Social Computing (SocialCom)*. 2011, 192–199.

[22] Correa, D., Sureka, A., and Sethi, R. WhACKY! - What Anyone Could Know About You From Twitter. In: *Proceedings of the 10th Annual International Conference on Privacy, Security and Trust (PST)*. 2012, 43–50.

[23] Cortis, K., Scerri, S., Rivera, I., and Handschuh, S. Discovering Semantic Equivalence of People Behind Online Profiles. In: *Proceedings of the 5th International Workshop on Resource Discovery (RED)*. 2012, 104–118.

[24] Diaz-Aviles, E. and Stewart, A. Tracking Twitter for Epidemic Intelligence: Case Study: EHEC/HUS Outbreak in Germany, 2011. In: *Proceedings of the 4th Annual ACM Web Science Conference (WebSci)*. 2012, 82–85.

[25] Dinur, I. and Nissim, K. Revealing Information While Preserving Privacy. In: *Proceedings of the 22nd ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems (PODS)*. 2003, 202–210.

[26] *Directive 95/46/EC of the European Parliament and of the Council on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data*. 1996.

[27] Domingos, P. and Richardson, M. Mining the network value of customers. In: *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2001, 57–66.

[28] Du, N., Song, L., Gomez-Rodriguez, M., and Zha, H. Scalable Influence Estimation in Continuous-time Diffusion Networks. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*. 2013, 3147–3155.

[29] Du, N., Song, L., Woo, H., and Zha, H. Uncover Topic-sensitive Information Diffusion Networks. In: *Artificial Intelligence and Statistics*. 2013, 229–237.

[30] Dwork, C. Differential Privacy: A Survey of Results. In: *Proceedings of the 5th International Conference on Theory and Applications of Models of Computation*. 2008, 1–19.

[31] Dwork, C., Kenthapadi, K., McSherry, F., Mironov, I., and Naor, M. Our Data, Ourselves: Privacy Via Distributed Noise Generation. In: *Advances in Cryptology (EUROCRYPT)*. 2006, 486–503.

[32] Dwork, C. and Naor, M. On the Difficulties of Disclosure Prevention in Statistical Databases or the Case for Differential Privacy. *Journal of Privacy and Confidentiality* 2, 1 (2008), 8.

[33] Endres, D. M. and Schindelin, J. E. A new Metric for Probability Distributions. *IEEE Transactions on Information Theory* 49, 7 (2003), 1858–1860.

[34] Erdos, P. and Rényi, A. On the Evolution of Random Graphs. *Publ. Math. Inst. Hung. Acad. Sci* 5, 1 (1960), 17–60.

[35] Eyharabide, V. and Amandi, A. Ontology-Based User Profile Learning. 36 (4 2012), 857–869.

[36] Feige, U. A Threshold of Ln N for Approximating Set Cover. *Journal of the ACM* 45, 4 (1998), 634–652.

[37] Feige, U., Mirrokni, V. S., and Vondrák, J. Maximizing Non-monotone Submodular Functions. *SIAM Journal of Computing* 40, 4 (2011), 1133–1153.

[38] Goga, O., Lei, H., Parthasarathi, S., Friedland, G., Sommer, R., and Teixeira, R. Exploiting Innocuous Activity for Correlating Users Across Sites. In: *WWW*. 2013.

[39] Goga, O., Loiseau, P., Sommer, R., Teixeira, R., and Gummadi, K. P. On the reliability of profile matching across large online social networks. In: *ACM KDD'15*.

[40] Goldenberg, J., Libai, B., and Muller, E. Using complex systems analysis to advance marketing theory development: Modeling heterogeneity effects on new product growth through stochastic cellular automata. *Academy of Marketing Science Review* 2001 (2001), 1.

[41] Gomez Rodriguez, M., Schölkopf, B., Pineau, L. J., et al. Submodular Inference of Diffusion Networks from Multiple Trees. In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*. 2012, 1–8.

[42] Gomez Rodriguez, M., Leskovec, J., and Schölkopf, B. Structure and Dynamics of Information Pathways in Online Media. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. WSDM '13. Rome, Italy, 2013, 23–32. ISBN: 978-1-4503-1869-3.

[43] Gomez-Rodriguez, M., Balduzzi, D., and Schoelkopf, B. Uncovering the Temporal Dynamics of Diffusion Networks. In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*. 2011.

[44] Gomez-Rodriguez, M., Leskovec, J., and Krause, A. Inferring Networks of Diffusion and Influence. In: *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2010, 1019–1028.

[45] Gomez-Rodriguez, M. and Schölkopf, B. Influence Maximization in Continuous Time Diffusion Networks. In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*. 2012, 313–320.

[46] Gomez-Rodriguez, M., Song, L., Du, N., Zha, H., and Schölkopf, B. Influence Estimation and Maximization in Continuous-Time Diffusion Networks. *ACM Transactions on Information Systems* 34, 2 (2016), 9:1–9:33.

[47] Granovetter, M. Threshold Models of Collective Behavior. *American Journal of Sociology* 83, 6 (1978), 1420–1443.

[48] Heatherly, R., Kantarcioglu, M., and Thuraisingham, B. Preventing Private Information Inference Attacks on Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 25, 8 (2013), 1849–1862.

[49] Hiemstra, D., Robertson, S., and Zaragoza, H. Parsimonious Language Models for Information Retrieval. In: *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 2004, 178–185.

[50] Iofciu, T., Fankhauser, P., Abel, F., and Bischoff, K. Identifying Users Across Social Tagging Systems. In: *ICWSM*. 2011.

[51] Irani, D., Webb, S., Li, K., and Pu, C. Large Online Social Footprints–An Emerging Threat. In: *Proceedings of the 2009 International Conference on Computational Science and Engineering - Volume 03 (CSE)*. 2009, 271–276.

[52] Irani, D., Webb, S., Li, K., and Pu, C. Modeling Unintended Personal-Information Leakage from Multiple Online Social Networks. *IEEE Internet Computing* 15, 3 (2011), 13–19.

[53] Iyer, R., Jegelka, S., and Bilmes, J. Fast Semidifferential-based Submodular Function Optimization. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning (ICML)*. 2013, III-855–III-863.

[54] Jones, R., Kumar, R., Pang, B., and Tomkins, A. "I Know What You Did Last Summer": Query Logs and User Privacy. In: *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management*. CIKM '07. 2007, 909–914.

[55] Kasivisiwanathan, S. P. and Smith, A. On the 'Semantics' of Differential Privacy: A Bayesian Formulation. *Journal of Privacy and Confidentiality* 6, 1 (2014), 1–16.

[56] Kasiviswanathan, S. P., Lee, H. K., Nissim, K., Raskhodnikova, S., and Smith, A. What can we learn privately? *SIAM Journal on Computing* 40, 3 (2011), 793–826.

[57] Kempe, D., Kleinberg, J., and Tardos, É. Maximizing the Spread of Influence Trough a Social Network. In: *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2003, 137–146.

[58] Khalil, E. B., Dilkina, B., and Song, L. Scalable Diffusion-aware Optimization of Network Topology. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2014, 1226–1235.

[59] Kifer, D. and Machanavajjhala, A. No Free Lunch in Data Privacy. In: *Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data*. 2011, 193–204.

[60] Koppel, M., Schler, J., and Argamon, S. Computational Methods in Authorship Attribution. *Journal of the American Society for Information Science and Technology* 60, 1 (2009), 9–26.

[61] Korula, N. and Lattanzi, S. An Efficient Reconciliation Algorithm for Social Networks. *Proceedings of the VLDB Endowment* 7, 5 (2014), 377–388.

[62] Kosinski, M., Bachrach, Y., Kohli, P., Stillwell, D., and Graepel, T. Manifestations of User Personality in Website Choice and Behaviour on Online Social Networks. *Machine Learning* 95, 3 (2014), 357–380.

[63] Krishnamurthy, B. and Wills, C. E. On the Leakage of Personally Identifiable Information via Online Social Networks. In: *Proceedings of the 2Nd ACM Workshop on Online Social Networks (WSON)*. 2009, 7–12.

[64] Labitzke, S., Taranu, I., and Hartenstein, H. What Your Friends Tell Others About You: Low Cost Linkability of Social Network Profiles. In: *SNA-KDD*. 2011.

[65] Lavrenko, V. and Croft, W. B. Relevance Based Language Models. In: *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2001, 120–127.

[66] Leskovec, J., Backstrom, L., and Kleinberg, J. Meme-tracking and the Dynamics of the News Cycle. In: *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2009, 497–506.

[67] Leskovec, J., Chakrabarti, D., Kleinberg, J., Faloutsos, C., and Ghahramani, Z. Kronecker Graphs: An Approach to Modeling Networks. *Journal of Machine Learning Research* 11 (2010), 985–1042.

[68] Leskovec, J., Lang, K. J., Dasgupta, A., and Mahoney, M. W. Statistical Properties of Community Structure in Large Social and Information Networks. In: *Proceedings of the 17th International Conference on World Wide Web (WWW)*. 2008, 695–704.

[69] Leskovec, J. and Sosič, R. SNAP: A General-Purpose Network Analysis and Graph-Mining Library. *ACM Transactions on Intelligent Systems and Technology (TIST)* 8, 1 (2016), 1.

[70] Li, N. and Li, T. t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. In: *Proceedings of the 23rd International Conference on Data Engineering (ICDE)*. 2007.

[71] Li, Y., Li, Y., Yan, Q., and Deng, R. H. Privacy Leakage Analysis in Online Social Networks. *Computers & Security* 49 (2015), 239–254.

[72] Liu, K. and Terzi, E. Towards Identity Anonymization on Graphs. In: *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data (SIGMOD)*. ACM, 2008, 93–106.

[73] Liu, Y., Gummadi, K. P., Krishnamurthy, B., and Mislove, A. Analyzing Facebook Privacy Settings: User Expectations vs. Reality. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*. 2011, 61–70.

[74] Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam, M. L-Diversity: Privacy Beyond K-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007).

[75] Maia, M., Almeida, J., and Almeida, V. Identifying User Behavior in Online Social Networks. In: *Proceedings of the 1st Workshop on Social Network Systems (SocialNets)*. 2008, 1–6.

[76] McCallister, E., Grance, T., and Scarfone, K. A. *SP 800-122. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII)*. Tech. rep. technical report. 2010.

[77] Mishari, M. A. and Tsudik, G. Exploring Linkability of User Reviews. In: *ESORICS*. 2012.

[78] Mondal, M., Druschel, P., Gummadi, K. P., and Mislove, A. Beyond Access Control: Managing Online Privacy via Exposure. In: *Proceedings of the Workshop on Useable Security*. 2014, 1–6.

[79] Mondal, M., Liu, Y., Viswanath, B., Gummadi, K. P., and Mislove, A. Understanding and Specifying Social Access Control Lists. In: *Proceedings of the 10th Symposium on Usable Privacy and Security (SOUPS)*. 2014, 11.

[80] Narayanan, A. and Shmatikov, V. Myths and Fallacies of "Personally Identifiable Information". *Communications of the ACM* 53, 6 (), 24–26.

[81] Narayanan, A. and Shmatikov, V. De-anonymizing Social Networks. In: *Proceedings of the 30th IEEE Symposium on Security and Privacy (S&P)*. 2009, 173–187.

[82] Nemhauser, G. L., Wolsey, L. A., and Fisher, M. L. An Analysis of Approximations for Maximizing Submodular Set Functions. *Mathematical Programming* 14, 1 (1978), 265–294.

[83] Northern, C. T. and Nelson, M. L. An Unsupervised Approach to Discovering and Disambiguating Social Media Profiles. In: *Proceedings of the 1st Mining Data Semantics Workshop*. 2011.

[84] Paul, M. J. and Dredze, M. You Are What You Tweet: Analyzing Twitter for Public Health. In: *Proceedings of the Fifth International Conference on Weblogs and Social Media*. 2011.

[85] Perito, D., Castelluccia, C., Ali Kâafar, M., and Manils, P. How Unique and Traceable Are Usernames? In: *Proceedings of the 11th International Symposium on Privacy Enhancing Technologies (PETS)*. 2011, 1–17.

[86] Pfitzmann, A. and Hansen, M. *A terminology for talking about privacy by data minimization: Anonymity, Unlinkability, Undetectability, Unobservability, Pseudonymity, and Identity Management*. `http://dud.inf.tu-dresden.de/literatur/Anon_Terminology_v0.34.pdf`. v0.34. 2010.

[87] Ponte, J. M. and Croft, W. B. A Language Modeling Approach to Information Retrieval. In: *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1998, 275–281.

[88] Proserpio, D., Goldberg, S., and McSherry, F. A Workflow for Differentially-private Graph Synthesis. In: *Proceedings of the 2012 ACM workshop on Workshop on Online Social Networks (WOSN)*. ACM, 2012, 13–18.

[89] Richardson, M. and Domingos, P. Mining Knowledge-sharing Sites for Viral Marketing. In: *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2002, 61–70.

[90] Rosenfeld, R. Two Decades of Statistical Language Modeling: Where do we go from Here? *Proceedings of the IEEE* 88, 8 (2000), 1270–1278.

[91] Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors. In: *Proceedings of the 19th International Conference on World Wide Web*. WWW '10. 2010, 851–860.

[92] Sala, A., Zhao, X., Wilson, C., Zheng, H., and Zhao, B. Y. Sharing Graphs using Differentially Private Graph Models. In: *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference (IMC)*. ACM, 2011, 81–98.

[93] Scanfeld, D., Scanfeld, V., and Larson, E. L. Dissemination of Health Information through Social Networks: Twitter and Antibiotics. *American Journal of Infection Control* 38, 3 (2010), 182–188.

[94] Scerri, S., Cortis, K., Rivera, I., and Handschuh, S. Knowledge Discovery in Distributed Social Web Sharing Activities. In: *Proceedings of the 3rd International Workshop on Modeling Social Media: Collective Intelligence in Social Media (MSM)*. 2012.

[95] Scerri, S., Gimenez, R., Herman, F., Bourimi, M., and Thiel, S. digital.me – Towards an Integrated Personal Information Sphere. In: *Federated Social Web Summit Europe*. 2011.

[96] Schelling, T. C. *Micromotives and macrobehavior*. WW Norton & Company, 2006.

[97] Seroussi, Y., Zukerman, I., and Bohnert, F. Authorship Attribution with Topic Models. *Computational Linguistics* 40, 2 (2014), 269–310.

[98] Sharma, N. K., Ghosh, S., Benevenuto, F., Ganguly, N., and Gummadi, K. Inferring Who-is-who in the Twitter Social Network. In: *Proceedings of the 2012 ACM Workshop on Workshop on Online Social Networks (WSON)*. 2012, 55–60.

[99] Sinha, A., Li, Y., and Bauer, L. What You Want is Not What You Get: Predicting Sharing Policies for Text-based Content on Facebook. In: *Proceedings of the 2013 ACM Workshop on Artificial Intelligence and Security (AISec)*. 2013, 13–24.

[100] Song, F. and Croft, W. B. A General Language Model for Information Retrieval. In: *Proceedings of the Eighth International Conference on Information and Knowledge Management (CIKM)*. 1999, 316–321.

[101] *Spokeo*. http://www.spokeo.com/.

[102] Svitkina, Z. and Fleischer, L. Submodular Approximation: Sampling-based Algorithms and Lower Bounds. *SIAM Journal on Computing* 40, 6 (2011), 1715–1737.

[103] Sweeney, L. K-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 5 (2002), 557–570.

[104] *The Online Social Network Reddit*. `http://www.reddit.com`. Accessed Nov 2014.

[105] Uzuner, Ö. and Katz, B. A Comparative Study of Language Models for Book and Author Recognition. In: *Natural Language Processing – IJCNLP 2005*. 2005, 969–980.

[106] Vinterbo, S. A. *A Note on the Hardness of the k-Ambiguity Problem*. Tech. rep. DSG-TR-2002-006. 2002.

[107] Wallinga, J. and Teunis, P. Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures. *American Journal of epidemiology* 160, 6 (2004), 509–516.

[108] Wang, C., Chen, W., and Wang, Y. Scalable Influence Maximization for Independent Cascade Model in Large-scale Social Networks. *Data Mining and Knowledge Discovery* 25, 3 (2012), 545–576.

[109] Wang, Y. and Wu, X. Preserving Differential Privacy in Degree-correlation based Graph Generation. *Transactions on Data Privacy* 6, 2 (2013), 127.

[110] Watts, D. J. and Dodds, P. S. Influentials, Networks, and Public Opinion Formation. *Journal of consumer research* 34, 4 (2007), 441–458.

[111] Xiao, Q., Chen, R., and Tan, K.-L. Differentially Private Network Data Release via Structural Inference. In: *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2014, 911–920.

[112] Xie, J., Knijnenburg, B. P., and Jin, H. Location Sharing Privacy Preference: Analysis and Personalized Recommendation. In: *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI)*. 2014, 189–198.

[113] Yao, Q., Shi, R., Zhou, C., Wang, P., and Guo, L. Topic-aware Social Influence Minimization. In: *Proceedings of the 24th International Conference on World Wide Web (WWW)*. 2015, 139–140.

[114] You, G.-w., Hwang, S.-w., Nie, Z., and Wen, J.-R. SocialSearch: Enhancing Entity Search with Social Network Matching. In: *EDBT/ICDT*. 2011.

[115] Zhai, C. Statistical Language Models for Information Retrieval A Critical Review. *Found. Trends Inf. Retr.* 2, 3 (2008), 137–213.

[116] Zhai, C. and Lafferty, J. A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems (TOIS)* 22, 2 (2004), 179–214.

[117] Zheleva, E. and Getoor, L. To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles. In: *Proceedings of the 18th International Conference on World Wide Web (WWW)*. 2009, 531–540.

[118] Zheleva, E. and Getoor, L. Privacy in Social Networks: A Survey. In: *Social Network Data Analytics*. 2011, 277–306.

[119] Zhou, B. and Pei, J. The k-Anonymity and l-Diversity Approaches for Privacy Preservation in Social Networks Against Neighborhood Attacks. *Knowledge and Information Systems* 28, 1 (2011), 47–77.