Community detection for access-control decisions: Analysing the role of homophily and information diffusion in Online Social Networks

Nicolás E. Díaz Ferreyra [a,b,*], Tobias Hecking [c], Esma Aïmeur [d], Maritta Heisel [b], H. Ulrich Hoppe [b]

[a] *Institute of Software Security, Hamburg University of Technology, 21079 Hamburg, Germany*
[b] *Department of Computer Science and Applied Cognitive Science, University of Duisburg–Essen, 47057 Duisburg, Germany*
[c] *Institute for Software Technology Intelligent and Distributed Systems, German Aerospace Center, 51147 Köln, Germany*
[d] *Department of Computer Science and Operations Research, University of Montréal, H3T 1J8 Montréal, Canada*

ARTICLE INFO

ABSTRACT

Access-Control Lists (ACLs) (a.k.a. "friend lists") are one of the most important privacy features of Online Social Networks (OSNs) as they allow users to restrict the audience of their publications. Nevertheless, creating and maintaining custom ACLs can introduce a high cognitive burden on average OSNs users since it normally requires assessing the trustworthiness of a large number of contacts. In principle, community detection algorithms can be leveraged to support the generation of ACLs by mapping a set of examples (i.e. contacts labelled as "untrusted") to the emerging communities inside the user's ego-network. However, unlike users' access-control preferences, traditional community-detection algorithms do not take the *homophily* characteristics of such communities into account (i.e. attributes shared among members). Consequently, this strategy may lead to inaccurate ACL configurations and privacy breaches under certain homophily scenarios. This work investigates the use of community-detection algorithms for the automatic generation of ACLs in OSNs. Particularly, it analyses the performance of the aforementioned approach under different homophily conditions through a simulation model. Furthermore, since private information may reach the scope of untrusted recipients through the re-sharing affordances of OSNs, information diffusion processes are also modelled and taken explicitly into account. Altogether, the removal of gatekeeper nodes is further explored as a strategy to counteract unwanted data dissemination.

1. Introduction

Online Social Networks (OSNs) like Twitter or Facebook are virtual spaces that allow people to connect with like-minded others by exchanging different types of media content including posts, links, and pictures [1]. To a large extent, social interaction inside these platforms resembles communication aspects of everyday life. Particularly, the exchange of private information, either inside or outside OSNs, is fundamental for creating and maintaining social relationships [2]. Hence, it is not surprising that people often disclose personal information in social media platforms in order to strengthen their bonds and maximize their social capital [3]. Nevertheless, keeping private information away from untrusted recipients becomes a challenging task for average users since OSNs place the members of disjoint social circles (e.g., family and work colleagues) under a same communication channel [3,4]. Consequently, this often leads to unintentional privacy breaches due to misalignments between the *intended* and the *actual* audience of online publications [5].

Access-Control Lists (ACLs) or just "friend lists" are one of the most salient privacy mechanisms of OSNs since they allow users to constrain the audience of the content they publish online [6,7]. Basically, ACLs are collections of contacts that are deemed *untrusted* by the user regarding the access to certain pieces of personal information. In principle, ACLs are an effective way to keep the information disclosed inside posts away from unintended recipients. For example, an ACL composed of work colleagues could be applied to restrict the visibility of a post with a negative comment about one's employer. However, creating custom ACLs can introduce a high cognitive burden on average users since it demands assessing the trustworthiness of a large number of contacts [8,9]. Furthermore, different ACLs must be created for different types of personal information since individuals' trustworthiness may vary from content to content [10]. Last but not least, keeping ACLs' internal consistency can demand a great effort given that users are likely to add or remove contacts from their network over time [11].
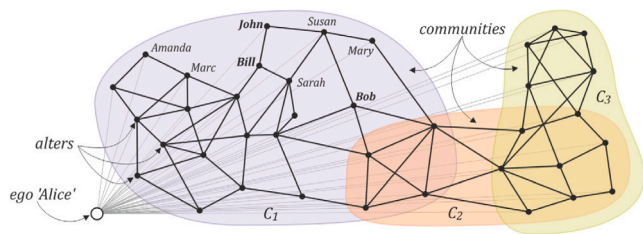

Fig. 1. Alice's ego-network. Untrusted friends John, Bill, and Bob are grouped under cluster C_1. An ACL containing all the nodes in C_1 is generated and recommended to Alice [17].

Access-Control Predictive Models (ACPMs) aim to reduce the burden of manual configurations through the automatic generation of custom ACLs [9]. Particularly, ACPMs leverage a set of *classification features* (e.g. personal attributes or network structure) to elaborate and recommend ACLs aligned with a set of *access-control preferences* provided by the user [7,12]. One approach consists of applying community detection algorithms for identifying clusters of untrusted members inside the user's ego-network (i.e. the network of connections between her friends) [12,13]. Under this approach, an ACL is created out of the members of the cluster that best fits the user's access-control preferences. In this case, preferences are specified through a set of contacts marked as "untrusted" by the user (Fig. 1). Such a community-based ACPM is suitable particularly in cases where accessing the personal attributes of network members (e.g. age, gender, workplace) is limited or not possible at all [13,14]. This is because traditional community-detection algorithms (such as Leading Eigenvector [15] or Multilevel Community [16]) can identify clusters without requiring information on the attribute values of its nodes (if any).

Motivation

Although community-based ACPMs only require structural network information for their application, they can nonetheless lead to inaccurate ACL configurations. Particularly, this is due to the influence of *homophily* in the formation of communities inside a network as on people's access-control preferences [18–20]. Basically, homophily refers to an organization principle in which the members of a network with similar characteristics tend to create connections with each other [21]. On the one hand, such a principle has a large impact on a network's structure and consequently in the formation of its communities [19]. Particularly, it is to expect that the members of an emerging cluster would hold a certain similarity with regard to a set of attribute values (e.g. location, gender or workplace) [22]. On the other hand, users' access-control preferences are also largely influenced by some of the attribute values shared among network members. For instance, in order to generate an ACL of work colleagues, a user would label as "untrusted" some of the contacts who share her same "workplace" attribute value. However, unlike the user's ACL preferences, traditional community-detection algorithms are not driven by attribute similarity causing flaws in the identification of untrusted communities [13,17, 23]. In particular, they may select best-fit clusters whose members do not portray similarities on the attribute values deemed relevant for the generation of the corresponding ACL [17]. Consequently, such a community-based ACPM may generate inaccurate ACL configurations and privacy vulnerabilities under certain homophily scenarios.

Another factor that can affect the efficacy of ACLs is information diffusion in OSNs. Particularly, OSNs are endowed with affordances that allow their users to re-share content from others making it available to an audience beyond the one defined by the content owner [24]. This can result in privacy breaches especially when a *trusted* member makes the user's post available to an *untrusted* one after re-sharing it [24,25]. Nevertheless, despite its importance in terms of privacy

and access-control, the dynamics of information diffusion in OSNs are often neglected by ACPMs including community-based ones [26]. Consequently, privacy violations can occur even when ACLs manage to properly identify all the untrusted recipients of a particular piece of private information. This calls for the elaboration of more robust solutions that are not only capable of satisfying users' access-control preferences, but also manage to prevent the propagation of personal information throughout the untrusted segments of a network.

Contribution

This work elaborates on the findings reported in [17] further investigating the impact of homophily and information diffusion in community-based ACPMs. Particularly, the performance of this approach is evaluated against different network topologies and information diffusion settings. For this, two simulation experiments are outlined, executed, and interpreted. In the first one, topologies are generated according to particular homophily conditions describing the attachment probability between nodes of similar characteristics. These network configurations are used thereafter to determine the precision, recall, and F1 score of community-detection algorithms when predicting ACLs. Unlike in [17], Links in Context [27], an attribute-based clustering approach, is also considered for performance evaluation along with traditional structure-based methods. In the second experiment, the dynamics of information diffusion (not considered in [17]) is introduced in the generated networks to elaborate countermeasures against unwanted data dissemination. In this case, the removal of gatekeeper nodes (i.e. trusted members linked to untrusted clusters) is investigated through a model of social influence as an approach for reducing the amount of personal information spread across untrusted segments of a network.

The rest of this paper is organized as follows. In the next section, related work on access-control prediction and information diffusion in OSNs is discussed and analysed. Following, Section 3 introduces the theoretical foundations of this paper. In particular, homophily-driven preferential attachment and social influence models are presented and described for later application. Sections 4 and 5 describe the methodology and results of simulation experiments 1 and 2, respectively. Such results together with their limitations are further discussed in Section 6. Finally, in Section 7, we outline the conclusions of this paper and introduce directions for future work.

2. Related work

Controlling the access to personal information is until now one of the main privacy challenges of OSNs. Particularly, such a challenge has motivated several contributions dealing with the automatic configuration and recommendation of ACLs. This section discusses related work elaborating on different ACPMs within the current literature. Likewise, advances on information diffusion and social influence are also analysed under the lens of access-control decisions.

2.1. Access-control prediction

Privacy and security scholars have developed several approaches to facilitate the definition of access-control policies in OSNs. Particularly, ACPMs seek to automate the generation of ACLs through machine learning [7,10,28–30], formal logic [31,32], and network analysis [14, 17,23] among other methods. On a large scale, ACPMs can be classified into *community-based* [14,17,23] or *attribute-based* [7,10,29,33], depending on whether they leverage communities or personal attributes for the automatic generation of access-control policies. For instance, Díaz Ferreyra et al. [29] introduced an attribute-based solution in which decision trees are generated to recommend adequate post audiences in OSNs. Under this approach, friends are classified into trusted or untrusted by applying a number of conditional tests over a set of

profile attributes such as age, gender, interest and education. A similar strategy is followed by Dong et al. [10] who elaborated on a classifier in which the sharing tendency between users together with the sensitivity of the content being disclosed are employed as audience predictors. Similar attributes along with demographic and location-related information were used by Ni et al. [7] in a machine learning solution that recommends personalized privacy policies for user-generated content in OSNs.

Despite their levels of accuracy, attribute-based solutions display some limitations related to the information on which access-control predictions are made. Particularly, deciding which attributes should be deemed as predictors is not a trivial decision. Furthermore, some attributes may not be available across different OSNs, making it difficult to engineer multi-platform solutions [14]. Conversely, community-based ACPMs do not suffer from these limitations and may be considered as less privacy invasive since their predictions are grounded on clustering algorithms that use solely the network structure as input. For example, Misra et al. [14] unveiled untrusted social circles in OSNs using the Clique Percolation Method, an approach which builds up communities out of fully-connected network sub-graphs. Certainly, hybrid solutions have also been introduced and discussed along with community and attribute-based ACPMs. Such is the case of Fangfang et al. [30] who elaborated a model for the automatic trust assessment of network members based on both, the connections and attribute similarities among them. Nevertheless, although hybrid solutions can show a high accuracy, their performance tends to decrease as the size of the network becomes larger [34]. Furthermore, a performance comparable to the one of hybrid ACPMs can be also achieved by community-based solutions within a shorter time frame [14,23].

2.2. Information diffusion in OSNs

A large body of research has been dedicated to the study of diffusion processes in complex networks [26,35–39]. Particularly, a significant number of works have explored approaches for the identification of the most influential users in a network with the aim of *maximizing* the spread of information [40]. For instance, Dhamal [35] elaborated on influence optimization through an adaptive seed strategy consisting of multiple phases of information diffusion. Under this method, influence maximization is achieved by selecting seed nodes at different stages according to the observed propagation of information over time. In line with this, Radion Purba et al. [36] introduced an information diffusion model that incorporates the engagement and activeness levels of Instagram users as indicators of their influence susceptibility and influence degree, respectively. Work concerning the identification of diffusion participants can also be found within the current literature. Such is the case of Li et al. [37] who developed a method for spotting members that are likely to forward viral information in Twitter. Particularly, they elaborated a ranking of potential hashtag adopters by analysing and forecasting the use of hashtags across multiple chains of followers.

All in all, influence maximization research is of great value for many real-world applications including online marketing [41], trending topic detection [42], and information summarization [43]. Nevertheless, information diffusion has also been analysed from an influence *minimization* perspective and applied in areas such as public health policies [44], misinformation [45], and cybersecurity [26,38,39]. For instance, Jia et al. [38] proposed limiting the propagation of adversarial content through the deletion of critical nodes and edges in a network. Likewise, Yan et al. [45] employed an edge-removal strategy to counteract the diffusion of rumours across OSNs. Overall, these applications share the common goal of stopping the spread of an agent (viruses, information, etc.) across the members of a network. However, access-control decisions in OSNs introduce the additional challenge of optimizing the utility of information disclosure. That is, maximizing the number of trusted recipients while reducing the number of untrusted ones. Prior work has introduced formal frameworks for analysing this

problem and algorithms for approximating its solution (cf. [26,39]). Still, to the best of our knowledge, such a challenge has not been tackled from a community-detection perspective. Furthermore, it has not been analysed yet under the lens of homophily.

3. Theoretical background

As discussed in Section 2.2, ACPMs often neglect the role of information diffusion when generating custom privacy policies. Moreover, homophily can also affect the performance of community-based approaches since it has a direct impact in the formation of communities inside OSNs. This work aims to investigate these aspects through simulation models of information diffusion and attributed scale-free networks. Such models and their main characteristics are introduced and discussed in the following subsections.

3.1. Homophilic preferential attachment

Large complex networks like the Internet or OSNs have caught the attention of researchers across different fields [46]. Empirical studies have shown that these networks share an important property: only a small number of nodes hold a big amount of connections to other nodes, whereas most nodes have just a few [47]. Networks containing such important nodes, or *hubs*, tend to be *scale-free* in the sense that the degree of these hubs (i.e. number of connections to other nodes) widely exceeds the average [47,48]. Up to now, scholars have proposed several evolution models for constructing scale-free networks [22,49–51]. Among those models, one of the most prominent ones is the one introduced by Barabasi and Albert [51]. This model indicates that two simple mechanisms, *growth* and *preferential attachment*, are responsible for the emergence of scale-free networks [47]. On the one hand, growth refers to the process in which at each time step a node with m ($\leq m_0$) links is added to the network and connected to m pre-existing nodes. Preferential attachment, on the other hand, describes the process by which new nodes prefer to link to the more connected nodes in the network (i.e. the hubs) [47].

In principle, the preferential attachment mechanism proposed by Barabasi and Albert does not take attribute similarity into consideration. Nevertheless, this approach has been enriched with homophily characteristics by other researchers resulting in more specific network evolution models [17,22,49]. For example, Kim et al. [22] introduced a *group-openness* mechanism for modelling the homophily and attachment probability between two nodes in a network. Under this approach, a node characterized with the attribute value s is considered a member of the group s. Then, the attachment probability between a node i of group s and a node j of group t is computed as a function of the openness factor Λ_s^t between the groups s and t. Such a factor can adopt values between 0 and 1 and indicates how closed (or open) is a node in group t to create links with other nodes in s [52]. Diaz Ferreyra et al. [17] adopted and extended this mechanism to situations in which nodes are characterized through multiple attribute values and therefore deemed members of more than one group. Particularly, homophily across different groups is expressed through an openness matrix Λ composed by the openness factors of all pairs of attribute values available in the network (see Appendix A for a detailed description). The corresponding preferential attachment model is adopted in this work for the generation of network topologies aligned with specific homophily conditions.

3.2. Independent cascade

Many efforts have been made to understand and recreate information diffusion processes in OSNs [53–55]. To a wide extent, such efforts have their origins in studies seeking to simulate the dynamics of epidemic diseases in biological networks. Overall, this is because information (just like a virus) begins to spread from a set of seed nodes
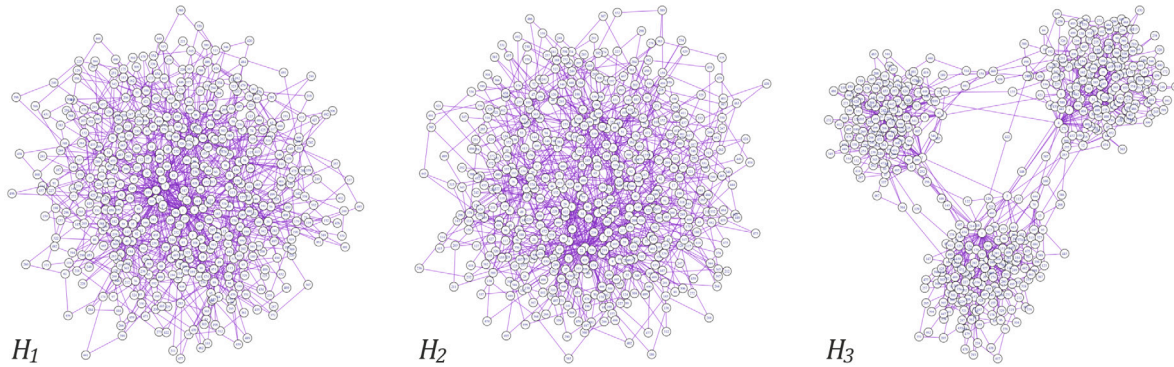
Fig. 2. Simulation output for homophily conditions H_1, H_2, and H_3.

to the rest of the network at a certain diffusion rate [54,55]. Different information diffusion models within the current literature have sought to recreate such a process making particular assumptions about its dynamics [54]. For example, progressive models like the Lineal Threshold (LT) [56] assume that, once infected, nodes cannot switch back to their original non-infected state. Conversely, non-progressive models like the Susceptible Infected Susceptible (SIS) [57] consider the scenario in which an infected node can return to its initial condition and therefore be infected many times. Particularly a progressive variation of this last one, the Susceptible Infected Recovered (SIR), has been applied in the context of the COVID-19 pandemic to simulate chains of contagion across communities [58].

Another approach widely applied to describe information diffusion processes in OSNs is the Independent Cascade (IC) [55,56]. Let us consider an ego-network representation $G = (V, E)$ where the node set V corresponds to the user's befriended contacts and E to the connections existing between them. The IC is a non-progressive model in which information spreads from an initial set of infected nodes $A_0 \subseteq V$ to the rest of the network members like a domino [54]. For the case of sharing a post in an OSN, A_0 could be defined as a sub-group of friends who notice the user's post once published and are likely to re-share it. Hence, at each time step $t \geq 0$ an infected node $v \in A_t$ can pass the information to an inactive (i.e. non-infected) neighbour $u : (u, v) \in E$ with probability p_{vu}. If successful, u becomes infected at time step $t + 1$, otherwise u stays inactive and v has no chance to infect it again. However, if u has more than one infected neighbour, it will be approached by each of them independently. Such a contagion process continues to unfold until there are no more infection attempts to be triggered. In Section 5.1, the IC model is adopted and instantiated to evaluate (i) the effectiveness of the predicted ACLs under information diffusion conditions, and (ii) elaborate countermeasures to prevent unwanted data dissemination in OSNs.

4. Simulation 1: Generation of ACLs

All in all, homophily is a process that may impact a network's structure and consequently impair the outcome of community-based ACPMs. This experiment aims to analyse the effect of homophily when generating ACLs from the emerging communities inside ego-networks. For this, such a community-based approach was put into practice in simulated network topologies under different homophily conditions. The following subsections describe the methodology employed in the experiment and the results obtained from its execution.

4.1. Methodology

In this experiment, the simulation approach introduced in [17] is followed. As it is shown in Fig. 3, a community-based ACPM requires (i) the user's ego-network and (ii) her privacy preferences in order
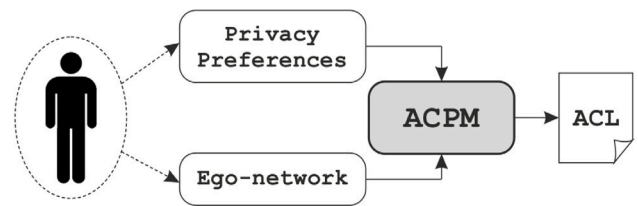

Fig. 3. Simulation approach for the generation of ACLs [17].

to generate a personalized ACL. Particularly, such preferences can be described through a small sample of untrusted friends. That is, some of those contacts who should be excluded from the audience of a certain piece of personal information. Therefore, this stage consists of a method for simulating ego-networks and a criterion for the selection of untrusted network members:

- *Simulation of ego-networks*: As discussed in Section 3.1, networks are generated through a homophily-driven preferential attachment model. In this case, the nodes of such networks are characterized with the attributes *gender*, *workplace*, and *location* as these are common to almost every OSN platform. Gender can adopt the values *male* or *female*, workplace can be *Starbucks*, *Google*, or *Ikea*, and location be either *Leeds*, or *York*. Under this model, the attachment probability between two nodes is computed through an openness matrix Λ describing the homophily conditions of the network. Basically, the values inside Λ can range from 0 to 1 and represent the strength of attribute similarity in the linking process (e.g. a value Λ_{York}^{Leeds} closer to zero describes a setting in which users located in *York* are less likely to connect with users living in *Leeds*). An extended description of this mechanism can be found in Appendix A.
- *Selection of untrusted nodes*: The selection of untrusted network members is guided by a hypothetical self-disclosure scenario in which a user working in *Ikea* (i.e. the *ego* of the simulated network) wishes to create an ACL to exclude her work colleagues from the audience of her publications. Hence, n nodes with the attribute value *workplace = Ikea* are selected from the generated ego-networks representing the user's access-control preferences. Particularly, nodes with the highest degree are selected as these are often the most influential ones in the network. The corresponding ACL is then built out of the community that brings together the largest amount of these untrusted nodes. For this, a community-detection algorithm (CDA) is employed to identify clusters of nodes inside the network under analysis.

The goal of this simulation experiment is to evaluate the performance of community-based ACLs under different homophily conditions.
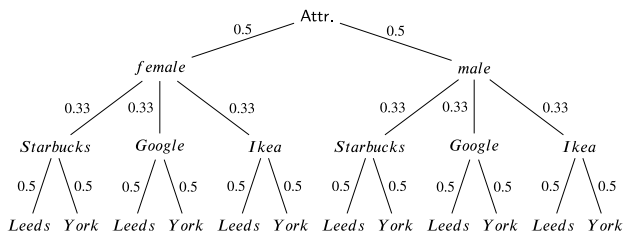
Fig. 4. Attributes probability distribution [17].

Table 1

Simulation 1: Results of configurations H_1, H_2, and H_3.

Conf.	CDA	Nº of clusters	ACL size	Precision	Recall	F1
H_1	MC	15	53	30.19%	8.94%	13.79%
	LE	28	49	48.98%	13.41%	21.05%
	LiC	234	272	36.03%	54.75%	43.46%
H_2	MC	15	67	50.75%	21.25%	29.96%
	LE	16	40	75.00%	18.75%	30.00%
	LiC	530	53	62.26%	26.63%	30.99%
H_3	MC	6	141	99.29%	81.87%	89.74%
	LE	6	172	97.67%	98.25%	97.96%
	LiC	32	175	97.14%	99.42%	98.27%

For this, a set of network topologies were generated out of different Λ values and used thereafter to produce the corresponding ACLs. Since *workplace* is the attribute that drives the selection of untrusted nodes, three Λ configurations were defined to simulate networks with a higher/lower degree of *workplace* homophily:

- H_1 : $\{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.7\}$
- H_2 : $\{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.3\}$
- H_3 : $\{\Lambda_{Google}^{Starbucks} = \Lambda_{Starbucks}^{Ikea} = \Lambda_{Google}^{Ikea} = 0.01\}$

As it can be observed, these configurations differ only in the values assigned to $\Lambda_{Google}^{Starbucks}$, $\Lambda_{Starbucks}^{Ikea}$ and Λ_{Google}^{Ikea} while the rest of the group-openness factors were set to one. These values were adopted from [17] as they lead to topologies with significant differences in structure and workplace homophily.

4.2. Execution and results

The simulation model described in the previous subsection was implemented using iGraph [59], a library for network analysis and visualization for R.[1] A total of three ego-networks of size $N = 500$ were simulated each aligned with the homophily conditions H_1, H_2 and H_3. In all cases, the attribute values of the nodes were assigned following the probability tree of Fig. 4. Likewise, $n = 10$ untrusted nodes were selected from each of the networks for the generation of the corresponding ACLs. Particularly, three different CDA were applied to unveil clusters inside the simulated networks: Leading Eigenvector (LE) [15], Multilevel Community (MC) [16], and Links in Context (LiC) [27]. Both LE and MC were initially assessed in a prior study (c.f. [17]). These methods rely solely on structural information to split the nodes of the network into a hierarchy of nested communities. Conversely, LiC also leverage the network's attributes to identify densely-connected groups of nodes [27]. Basically, LiC defines the context of each link in the network as the subset of attributes that are common to its endpoints. Thereby, it identifies communities of nodes through agglomerations of links sharing the same context. To the best of our knowledge, LiC has not been yet evaluated for the automatic generation of ACLs.

Fig. 2 illustrates the network topologies generated by the simulations. One can observe that, as the workplace openness becomes smaller (i.e. small $\Lambda_{Google}^{Starbucks}$, $\Lambda_{Starbucks}^{Ikea}$ and Λ_{Google}^{Ikea}), the nodes of the network tend to agglomerate around three big communities. Hence, the precision of the predicted ACLs (i.e. the percentage of untrusted nodes it contains) improves considerably from H_1 to H_2 and reaches its maximum in the homophily condition H_3 (Table 1). Particularly, the highest precision in H_1 and H_2 is achieved through the LE algorithm with a 48.98% and 75.00%, respectively. In the case of H_3 the best precision score corresponds to the algorithm MC with a value of 99.29%. On the other hand, the method's recall (i.e. percentage of untrusted nodes in the network included in the generated ACL) also reaches its maximum value in H_3 for all the clustering methods. However, only MC and LE improve their respective recall values from configurations

H_1 to H_2. Conversely, the recall of LiC drops from 54.7% in H_1 to 26.63% in H_2. The same phenomenon is observed for the F1 score (i.e the weighted average of precision and recall) which in the case of LiC decreases from 43.46% in H_1 to 30.99% in H_2. Nonetheless, LiC obtained the highest recall in H_1 and H_3 and the best F1 score across all homophily configurations.

In terms of execution time, MC and LE were able to identify clusters within seconds whereas LiC took more than 3 min in the best case (H_3).[2] With respect to the number of generated clusters, LiC unveiled more communities than LE and MC in all configurations. Particularly, LiC identified the largest number of clusters in H_2 (530 clusters) whereas the smaller amount corresponds to MC and LE in H_3 (6 clusters each). As for the ACL size, the average number of clustered members was 124.67 ± 127.61 in H_1, 53.33 ± 13.50 in H_2, and 162.67 ± 18.82 for the homophily condition H_3.

5. Simulation 2: Information diffusion

As discussed in Section 2.2, the re-sharing affordances of OSNs can impair the effectiveness of ACLs and violate, in turn, the privacy preferences of content-owners. This section introduces a simulation experiment for analysing the performance of community-based ACLs under different information diffusion settings. For this, the network topologies generated in Section 4 are infected using the IC model introduced in Section 3.2. Particularly, an ANOVA test is conducted to determine the effects of the network's homophily and the number of seed nodes on the percentage of infected nodes at the corresponding ACL. Furthermore, the effects of removing gatekeeper nodes are also evaluated and proposed as a means for counteracting unwanted data dissemination.

5.1. Methodology

Fig. 5 illustrates the main building blocks of the proposed simulation approach. First, a total of k gatekeeper nodes of the highest degree are removed from the ego-network under analysis. Such nodes are trusted members linked to one or more nodes inside the corresponding ACL and may therefore forward the information to untrusted network segments. Next, a group of s seed nodes is selected from outside the ACL and infected according to the premises of the IC model. Basically, this step recreates the situation in which the user's post reaches members of the trusted audience that are likely to re-share it. In principle, the IC model assumes that all infected nodes in the network can pass the information to their non-infected neighbours with the same probability. However, to a certain extent, the dynamics of information diffusion are related to homophily characteristics of OSNs [60]. Particularly, the flow of user-generated content across OSNs can be strongly influenced by attribute similarities among its users [61]. Therefore, the IC model was

---

[1] Link to the code repository: https://bit.ly/35kGmqZ.

[2] Values obtained with a 1.3 GHz dual-core processor and 8 GB 1867 MHz memory.

Table 2
Factorial ANOVA test.

Source	Sum Sq.	d.f.	Mean Sq.	F	p	η^2
Homophily	17 050.551	2	8525.276	77.622	0.000	0.080
Seeds	7988.157	1	7988.157	72.731	0.000	0.039
Rm-Gatekeepers	686 064.725	2	343 032.363	3123.266	0.000	0.778
Homophily * Seeds	1562.111	2	781.056	7.111	0.001	0.008
Homophily * Rm-Gatekeepers	43 990.027	4	10 997.507	100.131	0.000	0.184
Rm-Gatekeepers * Seeds	8761.669	2	438.335	3.991	0.019	0.004
Homophily * Seeds * Rm-Gatekeepers	884.828	4	221.207	2.014	0.090	0.005
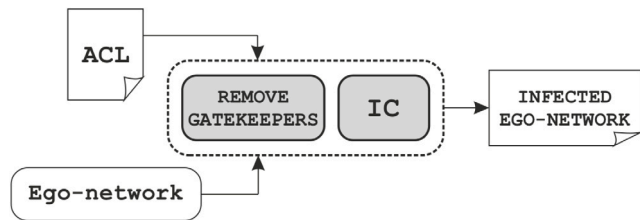
Fig. 5. Simulation approach for information diffusion.

Table 3
Tukey HSD Post-Hoc test.

Pair	Mean Diff.	p	95% CI
H_2 - H_1	6%	0.000	(4.58%, 7.42%)
H_3 - H_1	−0.96%	0.253	(−2.38%, 0.46%)
H_3 - H_2	−6.96%	0.000	(−8.37%, −5.54%)
$\frac{1}{3}$ Removed - 0 Removed	−30.95%	0.000	(−32.37%, −29.53%)
$\frac{2}{3}$ Removed - 0 Removed	−47.05%	0.000	(−48.46%, −45.63%)
$\frac{2}{3}$ Removed - $\frac{1}{3}$ Removed	−16.09%	0.000	(−17.51%, −14.67%)

adjusted so that the probability with which an infected node v spreads information to a non-infected neighbour u is computed as:

$$p_{vu} = \frac{\#Shared\,AttVal}{\#Total\,Att} \cdot \beta$$

where $\#Total\,Att$ corresponds to the number of attributes characterizing the network (3 in this case) and $\#Shared\,AttVal$ to the number of attribute values shared between v and u. The parameter β represents the maximum infection probability among all network members and can adopt values ranging between 0 and 1.

5.2. Execution and results

The effects of homophily along with the number of seed nodes and removed gatekeeper nodes were analysed in a $3 \times 2 \times 3$ factorial experimental design. Particularly, $k = 0$, $\frac{1}{3}$ or $\frac{2}{3}$ of the getaway nodes were removed from the networks corresponding to the homophily configurations H_1, H_2 and H_3. In addition, these networks were infected with $s = 75$ or 150 seed nodes as described in Section 5.1. In all cases, the β parameter was set to 0.6 to represent a tendency of the network members towards information re-sharing. Likewise, all ACLs were generated using the LE algorithm since it was the one with the best overall precision in Simulation 1. The IC model was executed 100 times for each of the $3 \times 2 \times 3 = 18$ experimental conditions resulting in 1800 simulation runs.

An ANOVA test was conducted on the simulation results to determine the effects of (i) homophily, (ii) diffusion seeds, and (iii) removed gatekeepers on the percentage of infected ACL members. At first, a Levene's test determined that the dataset resulting from the simulation outputs did not meet ANOVA's assumption of homogeneity of variance. However, this condition can be relaxed for large samples ($N \geq 30$) and when the number of observations is equally distributed (or nearly so) across the different factors and levels [62]. Since these conditions are met by the generated data set (i.e. N = 1800 and 100 observations per level), the ANOVA analysis was carried out accordingly. As summarized in Table 2, all of the main factors were significant for the percentage of infected ACL members after the execution of the IC model ($p < 0.001$). Particularly *homophily* yielded an effect size of $\eta^2 = 0.080$, indicating that around 8% of the variance in the percentage of infected ACL nodes was explained by this factor ($F_{2,1782} = 77.622, p = 0.000$). Such a variance is affected in 3.9% by the *diffusion seeds* ($F_{1,1782} = 72.731, p = 0.000$) and in 77.8% by the *removed gatekeepers* ($F_{2,1782} = 3123.266, p = 0.000$). A significant interaction effect of 0.008 was observed between *homophily* and *diffusion seeds* ($F_{2,1782} = 7.111, p = 0.001$). Likewise, a

significant effect of 0.184 was yielded from the interaction between *homophily* and *removed gatekeepers* ($F_{4,1782} = 100.131, p = 0.000$). However, no significant effects were observed from the interaction of all three factors altogether ($F_{4,1782} = 2.014, p = 0.090$).

In order to further examine the significant main effects, a Tukey HSD Post-Hoc test was conducted (Table 3). Particularly, an average drop of 30.95% in the infected ACL nodes was observed after removing $\frac{1}{3}$ of the gatekeeper nodes and 47.05% after removing $\frac{2}{3}$ ($p < 0.001$). Likewise, the average difference of −16.10% between the $\frac{2}{3}$ and $\frac{1}{3}$ gatekeeper removal conditions was also found significant ($p < 0.001$). On the other hand, the percentage of infected ACL members increases 6% on average from homophily condition H_1 to H_2 and decreases about 6.96% from H_2 to H_3 ($p < 0.001$). However, no significant differences were observed between homophily conditions H_3 and H_1.

6. Discussion

Several remarks can be drawn upon the results obtained from the simulation experiments. On the one hand, the performance of community-based ACPMs is closely related to the homophily of the network under analysis and the attribute values that are critical for the ACL configuration. This can be observed from the results of Simulation 1 where the precision of the predicted ACL becomes higher as the homophily in workplace increases. Particularly, LE resulted the method with the highest overall precision making it more suitable than MC and LiC for cases in which the costs of false positives (i.e. classifying trusted members as untrusted) are deemed high. Conversely, the recall and F1 values obtained in this experiment suggest that LiC would qualify better for cases where the penalty of false negatives (i.e. classifying untrusted members as trusted) is larger. In part, this is explained by the number and size of the communities generated by this method which, compared to MC and LE, result much larger. Nonetheless, the recall and F1 scores of LE are close to the ones of LiC in both, H_2 and H_3, making it also a good approach in such a case with the advantage of being computationally more effective.

Overall, the F1 values of LiC show that its loss in terms of precision is not so large. However, this metric along with the method's recall did not improve from configuration H_1 to H_2 (something that did happen with MC and LE). Furthermore, LiC produced much more false positives than MC or LE making it less suitable for cases in which Type 1 errors are considered critical. In part, this may be related to the approach followed by this method when partitioning the network space into overlapping communities. Particularly, LiC determines the optimal cut level of the resulting dendrogram through a measure of its
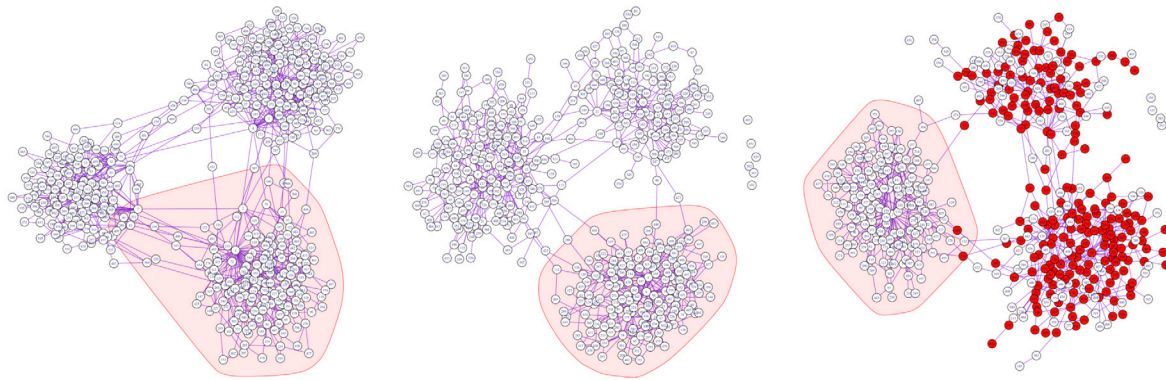
Fig. 6. Infection sequence for H_3: ACL (left), after removing 2/3 of gatekeepers (centre), after information diffusion with 150 seeds (right).

partition density which tends to divide the network into many small communities. Moreover, such communities may in turn be included in larger ones which is not the case for LE nor MC. This has a direct impact in the detection of the community that best fits the user's privacy preferences in terms of size and composition. Hence, such an aspect should be further investigated and adjusted so the LiC algorithm becomes more suitable for access-control prediction.

On the other hand, the results of Simulation 2 suggest that homophily can impact significantly the performance of community-based ACLs under information diffusion conditions. Likewise, such a performance is also affected by the number of seed nodes spreading the information across the network at $t = 0$. Nevertheless, the largest effect size and the highest drop in the percentage of infected ACL members is given by the number of gatekeeper nodes removed from the network. Certainly, it is to expect that as more gatekeeper nodes get removed, fewer infections inside the corresponding ACL will be observed. However, this has also a negative impact on the trusted audience's extent since gatekeepers are, after all, nothing but trusted network members. Furthermore, as observed in Fig. 6 some trusted nodes get disconnected from the network as a consequence of applying such a removal strategy. Hence, such a collateral damage should be minimized so the influence range of the content being disclosed is not significantly reduced. Particularly, these side effects should be taken explicitly into consideration for the elaboration of an adequate gatekeeper removal criterion.

To a certain extent, the results yielded by the simulation experiments of this work are subject to limitations. Particularly, such results may be affected by the set-up parameters of the simulation models as well as by the size and attribute characterization of the generated networks. On the one hand, the average size of ego-networks can largely differ from one OSN to another. For instance, it was estimated that around 38.35% of Facebook users in the United States had between 500 and 200 friends by 2016.[3] However, this number can easily reach the order of thousands in other OSNs like Instagram.[4] On the other hand, the type of attributes proposed in Section 4.2 as well as the distribution of their values may not necessarily represent actual OSNs conditions. Instead, it is to expect that certain attribute values will prevail over others and may not be equally distributed across the network members [63]. Finally, the probability of influence among OSNs users is subject to factors that go beyond attribute similarity. For instance, it can be affected by individual relations, time, and network centrality [40]. Hence, further experimental studies should closely consider these factors in order to characterize the IC model more adequately.

7. Conclusion and future work

The experimental results of this work provide valuable insights into the role of homophily and information diffusion when applying community-based ACPMs. Particularly homophily was shown to be a critical aspect for the generation of adequate ACL configurations through community-detection methods. In this sense, it is of paramount importance to have a good understanding of the homophily conditions at the network under analysis before applying clustering methods for ACL prediction. On the other hand, avoiding unwanted data dissemination remains an open challenge for access-control policies in OSNs. In principle, this issue can be mitigated by excluding a percentage of gatekeeper nodes from the members of the trusted audience. Nevertheless, side effects on the utility of information disclosure should be taken into consideration when applying such a countermeasure. For this, a removal criterion of gatekeeper nodes should seek to maximize the influence of the shared information while minimizing the chances of its propagation towards untrusted network segments.

There are multiple questions and future research directions that arise from the results of this work. One of them corresponds to conducting a study of community-based ACPMs under actual network conditions. This includes an analysis of attribute similarity and influence processes in OSNs for an adequate characterization of the simulation models. Likewise, evaluating other community detection approaches for access-control prediction will be a matter of future investigations. On the other hand, we expect to explore the models and principles presented in this work on areas outside privacy and security. Particularly the removal of gatekeeper nodes under certain homophily conditions could be applied to the design of public health policies, for instance, in elaborating social distancing strategies to control the spread of the COVID-19 pandemic. In such a case, more specific models for the simulation of epidemic diseases will be investigated and evaluated along with the principles of gatekeeper removal discussed throughout this work.

CRediT authorship contribution statement

Nicolás E. Díaz Ferreyra: Conceptualization, Methodology, Investigation, Writing – original draft, Formal analysis, Data curation. **Tobias Hecking:** Methodology, Investigation, Validation, Writing – review & editing. **Esma Aïmeur:** Validation, Writing – review & editing. **Maritta Heisel:** Validation, Writing – review & editing. **H. Ulrich Hoppe:** Validation, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

---

[3] Statista: Average number of Facebook friends of users in the United States in 2016 - https://bit.ly/3oljuyY.

[4] Statista: Worldwide Instagram follower growth rate from January to June 2019, by profile size - https://bit.ly/2JPlxwo.

Acknowledgements

Appendix A. Group openness

As introduced in Section 3.1, Kim et al. [22] defined a *group-openness* mechanism to investigate the role of homophily in the evolution of scale-free networks. Particularly, such approach considers that nodes sharing a particular attribute value s are members of a common group s. Then, the group-openness factor Λ_s^t between two groups s and t is defined as:

$$\Lambda_s^t = \begin{cases} \Lambda, & \text{if } s \neq t \\ 1, & \text{if } s = t \end{cases} \qquad 0 \leq \Lambda \leq 1 \tag{1}$$

where the homophily index Λ is a real number between 0 and 1 [52]. Particularly a value of $\Lambda = 0$ describes the case in which nodes in s are completely reluctant to create ties with others who are members of group t. Conversely, for $\Lambda = 1$, members of s connect openly with other nodes outside their group. Such Λ values can be employed to define how strong (or weak) is the role of homophily between two nodes in terms of group membership.

$$\mathbf{\Lambda} = \begin{bmatrix} \Lambda_{male}^{male} & \Lambda_{male}^{female} & \cdots & \Lambda_{male}^{York} \\ \Lambda_{female}^{male} & \Lambda_{female}^{female} & \cdots & \Lambda_{female}^{York} \\ \vdots & \vdots & \ddots & \vdots \\ \Lambda_{York}^{male} & \Lambda_{York}^{female} & \cdots & \Lambda_{York}^{York} \end{bmatrix} = \begin{bmatrix} 1 & 0.7 & \cdots & 0.3 \\ 0.7 & 1 & \cdots & 0.5 \\ \vdots & \vdots & \ddots & \vdots \\ 0.3 & 0.5 & \cdots & 1 \end{bmatrix}$$

Fig. 7. Group-openness matrix.

As mentioned in Section 4, the nodes of the networks analysed in this work are characterized with the attributes gender, workplace and location. Particularly, a node can adopt the values *male* or *female* for the "gender" attribute, *Starbucks*, *Google*, or *Ikea* for "workplace", and *York* or *Leeds* for "location". Therefore, each network consists of 7 groups making it possible to define up to $C_{7,2} = \frac{7!}{2!(7-2)!} = 21$ group openness factors. For instance, Λ_{York}^{Leeds} can be used to specify how open (or closed) is people from *York* to connect with others from *Leeds*. Likewise, Λ_{Google}^{Ikea} can describe how likely are workers from *Google* to create links with others from *Ikea*. In sum, the information concerning all homophily factors of a network can be expressed through a group-openness matrix $\Lambda_{7 \times 7}$ as shown in Fig. 7. Consequently, the total homophily factor between a node i and another node j can be defined as:

$$\mathcal{H}_P^Q = \prod_{\substack{p \in P \\ q \in Q}} \Lambda_{p,q} \qquad \begin{array}{l} P = \text{groups to which node } i \text{ belongs} \\ Q = \text{groups to which node } j \text{ belongs} \end{array} \tag{2}$$

where P and Q are the group sets to which i and j belong, respectively.

Appendix B. Preferential attachment model

In the Barabasi and Albert model, a node's probability of creating new connections with others depends exclusively on its degree. However, Eq. (2) can be used to introduce the role of attribute similarity

in the estimation of the attachment probability between network members. Consequently, the probability Π_i^{PQ} that a new node of group-set Q is linked to node i of group-set P is defined as:

$$\Pi_i^{PQ} = \frac{k_i^P . \mathcal{H}_P^Q}{\sum_j k_j^M . \mathcal{H}_M^Q} \qquad \begin{array}{l} k_i^P = \text{degree of node } i \text{ from} \\ \text{group-set } P \\ \mathcal{H}_P^Q = \text{total homophily factor be-} \\ \text{tween group-set } P \text{ and group-set } Q \end{array} \tag{3}$$

where k_i^P is the degree of node i from group-set P, and $\mathcal{H}_P^Q$ represents the total homophily factor between the group-set of the new node Q and the group-set of node i. This attachment rule describes the case in which a node without any connections is incorporated to the network. Nevertheless, new links may also emerge between existing network members over time. Particularly, such a probability Π_{ij}^{PQ} that an existing node i links to another existing one j is defined as:

$$\Pi_{ij}^{PQ} = \frac{k_i^P . k_j^Q . \mathcal{H}_P^Q}{\sum_l \sum_{m>l} k_l^M . k_m^N . \mathcal{H}_M^N} \tag{4}$$

where k_i^P and k_j^Q are the degrees of node i and of node j respectively and $\mathcal{H}_P^Q$ is the total homophily factor between their corresponding group-sets P and Q. This attachment rule together with the one of Eq. (3) were used to generate the network topologies analysed in this paper.

Appendix C. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.osnem.2022.100203.

References

[1] J. Penni, The future of online social networks (OSN): A measurement analysis using social media tools and application, Telemat. Inform. 34 (5) (2017) 498–517.

[2] Y.-C. Wang, M. Burke, R. Kraut, Modeling self-disclosure in social networking sites, in: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, in: CSCW '16, Association for Computing Machinery, New York, NY, USA, 2016, pp. 74–85, http://dx.doi.org/10.1145/2818048.2820010.

[3] N.C. Krämer, J. Schäwel, Mastering the challenge of balancing self-disclosure and privacy in social media, Curr. Opin. Psychol. 31 (2020).

[4] J. Vitak, The impact of context collapse and privacy on social network site disclosures, J. Broadcast. Electron. Media 56 (4) (2012) 451–470.

[5] H. Schwartz-Chassidim, O. Ayalon, T. Mendel, R. Hirschprung, E. Toch, Selectivity in posting on social networks: The role of privacy concerns, social capital, and technical literacy, Heliyon 6 (2) (2020) e03298.

[6] R. Hirschprung, E. Toch, H. Schwartz-Chassidim, T. Mendel, O. Maimon, Analyzing and optimizing access control choice architectures in online social networks, ACM Trans. Intell. Syst. Technol. (TIST) 8 (4) (2017) 1–22.

[7] M. Ni, Y. Zhang, W. Han, J. Pang, An empirical study on user access control in online social networks, in: Proceedings of the 21st ACM on Symposium on Access Control Models and Technologies, 2016, pp. 13–23.

[8] M. Namara, H. Sloan, P. Jaiswal, B.P. Knijnenburg, The potential for user-tailored privacy on facebook, in: 2018 IEEE Symposium on Privacy-Aware Computing (PAC), IEEE, 2018, pp. 31–42.

[9] J. Ahmed, S. Villata, G. Governatori, Information and friend segregation for online social networks: A user study, Ai Soc. 34 (4) (2019) 753–766.

[10] C. Dong, H. Jin, B.P. Knijnenburg, Ppm: A privacy prediction model for online social networks, in: International Conference on Social Informatics, Springer, 2016, pp. 400–420.

[11] M. Mondal, G.S. Yilmaz, N. Hirsch, M.T. Khan, M. Tang, C. Tran, C. Kanich, B. Ur, E. Zheleva, Moving beyond set-it-and-forget-it privacy settings on social media, in: Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 991–1008.

[12] T. Khazaei, L. Xiao, R.E. Mercer, A. Khan, Detecting privacy preferences from online social footprints: A literature review, in: IConference 2016 Proceedings, 2016.

[13] G. Misra, J.M. Such, H. Balogun, Non-sharing communities? An empirical study of community detection for access control decisions, in: 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE, 2016, pp. 49–56.

[14] G. Misra, J.M. Such, React: Recommending access control decisions to social media users, in: Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017, 2017, pp. 421–426.

[15] M.E.J. Newman, Finding community structure in networks using the eigenvectors of matrices, Phys. Rev. E 74 (2006) 036104, http://dx.doi.org/10.1103/PhysRevE.74.036104.

[16] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks, J. Statist. Mech.: Theory Exp. 2008 (10) (2008).

[17] N.E. Díaz Ferreyra, T. Hecking, H.U. Hoppe, M. Heisel, Access-control prediction in social network sites: Examining the role of homophily, in: Social Informatics, Springer International Publishing, Cham, 2018, pp. 61–74.

[18] G. Towhidi, A.P. Sinha, M. Srite, H. Zhao, Trust decision-making in online social communities: A network-based model, J. Comput. Inf. Syst. (2020) 1–11.

[19] A. De Salve, B. Guidi, L. Ricci, P. Mori, Discovering homophily in online social networks, Mob. Netw. Appl. 23 (6) (2018) 1715–1726.

[20] B. Gao, B. Berendt, Friends and circles—A design study for contact management in egocentric online social networks, in: Online Social Media Analysis and Visualization, Springer, 2014, pp. 129–161.

[21] M. McPherson, L. Smith-Lovin, J.M. Cook, Birds of a feather: Homophily in social networks, Annu. Rev. Sociol. 27 (1) (2001) 415–444.

[22] K. Kim, J. Altmann, Effect of homophily on network formation, Commun. Nonlinear Sci. Numer. Simul. 44 (2017) 482–494.

[23] B. Gao, B. Berendt, Circles, posts and privacy in egocentric social networks: An exploratory visualization approach, in: Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, 2013, pp. 792–796.

[24] M. Humbert, B. Trubert, K. Huguenin, A survey on interdependent privacy, ACM Comput. Surv. 52 (6) (2019) 1–40.

[25] L. Yu, S.M. Motipalli, D. Lee, P. Liu, H. Xu, Q. Liu, J. Tan, B. Luo, My friend leaks my privacy: Modeling and analyzing privacy in social networks, in: Proceedings of the 23nd ACM on Symposium on Access Control Models and Technologies, 2018, pp. 93–104.

[26] M. Backes, M. Gomez-Rodriguez, P. Manoharan, B. Surma, Reconciling privacy and utility in continuous-time diffusion networks, in: 2017 IEEE 30th Computer Security Foundations Symposium (CSF), IEEE, 2017, pp. 292–304.

[27] T. Hecking, H.U. Hoppe, Links in context: Detecting and describing the nested structure of communities in node-attributed networks, in: International Conference on Complex Networks and their Applications, Springer, 2018, pp. 195–206.

[28] G. Misra, J.M. Such, PACMAN: PErsonal agent for access control in social media, IEEE Internet Comput. 21 (6) (2017) 18–26.

[29] N. Díaz Ferreyra, R. Meis, M. Heisel, At your own risk: Shaping privacy heuristics for online self-disclosure, in: 2018 16th Annual Conference on Privacy, Security and Trust (PST), 2018, pp. 1–10.

[30] F. Shan, J. Liu, X. Wang, W. Liu, B. Zhou, A smart access control method for online social networks based on support vector machine, IEEE Access 8 (2020) 11096–11103.

[31] Y. Rafiq, L. Dickens, A. Russo, A.K. Bandara, M. Yang, A. Stuart, M. Levine, G. Calikli, B.A. Price, B. Nuseibeh, Learning to share: Engineering adaptive decision-support for online social networks, in: Proceedings of the 32nd IEEE/ACM International Conference on Automated Software Engineering (ASE), IEEE, 2017, pp. 280–285.

[32] S. Vahabli, R. Ravanmehr, A novel trust-based access control for social networks using fuzzy systems, World Wide Web 22 (6) (2019) 2241–2265.

[33] A. Squicciarini, S. Karumanchi, D. Lin, N. Desisto, Identifying hidden social circles for advanced privacy configuration, Comput. Secur. 41 (2014) 40–51.

[34] S. Ding, Z. Yue, S. Yang, F. Niu, Y. Zhang, A novel trust model based overlapping community detection algorithm for social networks, IEEE Trans. Knowl. Data Eng. (2019).

[35] S. Dhamal, Effectiveness of diffusing information through a social network in multiple phases, in: 2018 IEEE Global Communications Conference (GLOBECOM), IEEE, 2018, pp. 1–7.

[36] K.R. Purba, D. Asirvatham, R.K. Murugesan, Influence maximization diffusion models based on engagement and activeness on instagram, J. King Saud Univ.-Comput. Inf. Sci. (2020).

[37] C.-T. Li, Y.-J. Lin, M.-Y. Yeh, Forecasting participants of information diffusion on social networks with its applications, Inform. Sci. 422 (2018) 432–446.

[38] F. Jia, K. Zhou, C. Kamhoua, Y. Vorobeychik, Blocking adversarial influence in social networks, in: Proceedings of the 11th Conference on Decision and Game Theory for Security (GAMESEC), 2020.

[39] D. Gorla, F. Granese, C. Palamidessi, Enhanced models for privacy and utility in continuous-time diffusion networks, in: International Colloquium on Theoretical Aspects of Computing, Springer, 2019, pp. 313–331.

[40] Y. Li, J. Fan, J. Wang, K.-L. Tan, Influence maximization on social graphs: A survey, IEEE Trans. Knowl. Data Eng. 30 (10) (2018) 1852–1872.

[41] Q. Zeng, M. Zhong, Y. Zhu, J. Li, Business location selection based on geo-social networks, in: International Conference on Database Systems for Advanced Applications, Springer, 2020, pp. 36–52.

[42] Z. Miao, A. Chen, Y. Fang, J. He, Y. Zhou, W. Zhang, H. Zha, Cost-effective online trending topic detection and popularity prediction in microblogging, ACM Trans. Inf. Syst. (TOIS) 35 (3) (2016) 1–36.

[43] L. Shi, H. Tong, J. Tang, C. Lin, Vegas: Visual influence graph summarization on citation networks, IEEE Trans. Knowl. Data Eng. 27 (12) (2015) 3417–3431.

[44] Y. Zhang, A. Ramanathan, A. Vullikanti, L. Pullum, B.A. Prakash, Data-driven efficient network and surveillance-based immunization, Knowl. Inf. Syst. 61 (3) (2019) 1667–1693.

[45] R. Yan, Y. Li, W. Wu, D. Li, Y. Wang, Rumor blocking through online link deletion on social networks, ACM Trans. Knowl. Discov. Data (TKDD) 13 (2) (2019) 1–26.

[46] A.-L. Barabási, Scale-free networks: A decade and beyond, Science 325 (5939) (2009) 412–413.

[47] A.-L. Barabási, Network Science, Cambridge University Press, 2016.

[48] J. Jiang, S. Wen, S. Yu, Y. Xiang, W. Zhou, H. Hassan, The structure of communities in scale-free networks, Concurr. Comput.: Pract. Exper. 29 (14) (2017) e4040.

[49] J. Khan, S. Lee, Online social networks evolution model based on homophily and preferential attachment, Symmetry 10 (11) (2018) 654.

[50] C. Dangalchev, Generation models for scale-free networks, Physica A 338 (3–4) (2004) 659–671.

[51] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (5439) (1999) 509–512, http://dx.doi.org/10.1126/science.286.5439.509, arXiv:http://science.sciencemag.org/content/286/5439/509.full.pdf URL: http://science.sciencemag.org/content/286/5439/509.

[52] K. Kim, J. Altmann, J. Hwang, The impact of the subgroup structure on the evolution of networks: An economic model of network evolution, in: 2010 INFOCOM IEEE Conference on Computer Communications Workshops, 2010, pp. 1–9, http://dx.doi.org/10.1109/infocomw.2010.5466705.

[53] M. Li, X. Wang, K. Gao, S. Zhang, A survey on information diffusion in online social networks: Models and methods, Information 8 (4) (2017) 118.

[54] B. Chang, T. Xu, Q. Liu, E.-H. Chen, Study on information diffusion analysis in social networks and its applications, Int. J. Autom. Comput. 15 (4) (2018) 377–401.

[55] K. Li, L. Zhang, H. Huang, Social influence analysis: Models, methods, and evaluation, Engineering 4 (1) (2018) 40–46.

[56] P. Shakarian, A. Bhatnagar, A. Aleali, E. Shaabani, R. Guo, The independent cascade and linear threshold models, in: Diffusion in Social Networks, Springer, 2015, pp. 35–48.

[57] M.E. Newman, The structure and function of complex networks, SIAM Rev. 45 (2) (2003) 167–256.

[58] I. Cooper, A. Mondal, C.G. Antonopoulos, A SIR model assumption for the spread of COVID-19 in different communities, Chaos Solitons Fractals 139 (2020) 110057.

[59] G. Csardi, T. Nepusz, The igraph software package for complex network research, InterJournal Complex Systems (2006) 1695, URL: http://igraph.org.

[60] M. De Choudhury, H. Sundaram, A. John, D.D. Seligmann, Analyzing the dynamics of communication in online social networks, in: B. Furht (Ed.), Handbook of Social Network Technologies and Applications, Springer US, Boston, MA, 2010, pp. 59–94, http://dx.doi.org/10.1007/978-1-4419-7142-5_4.

[61] Z. Qin, Z. You, H. Jin, X. Gan, J. Wang, Homophily-driven evolution increases the diffusion accuracy in social networks, IEEE Trans. Netw. Sci. Eng. (2020).

[62] S.F. Sawyer, Analysis of variance: The fundamental concepts, J. Man. Manip. Ther. 17 (2) (2009) 27E–38E.

[63] Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, A. Kaltenbrunner, The length of bridge ties: Structural and geographic properties of online social interactions, in: Proceedings of the 6th International AAAI Conference on Weblogs and Social Media, 2012.