PELTE: Privacy Estimation of Images from Tags

Extended Abstract

Abdurrahman Can Kurtan
Department of Computer Engineering
Bogazici University
Istanbul, Turkey
can.kurtan@boun.edu.tr

Pınar Yolum
Department of Information and Computing Sciences
Utrecht University
Utrecht, The Netherlands
p.yolum@uu.nl

ABSTRACT

Image sharing is a service offered by many online social networks. In order to preserve privacy of images, users need to think through and set the privacy settings for each image that they upload. This is difficult for two main reasons: First, research shows that many times users do not know their own privacy preferences, but only become aware of them over time. Second, even when users know their privacy preferences, specifying these policies is cumbersome and requires too much effort, interfering with the quick sharing behavior expected on an social network. Accordingly, this paper proposes an agent-based approach, PELTE, that predicts the privacy setting of images using their content tags. Each user agent makes use of the privacy settings that its user have set for previous images to predict the privacy setting for a new uploaded one automatically. When in doubt, the agent analyzes the sharing behavior of other trusted agents to make a recommendation to its user about what is private. Contrary to existing approaches that assume a centralized online social network, our approach is distributed and thus each agent can only view the privacy settings of the images that it has shared or those that have been shared with it.

ACM Reference Format:
Abdurrahman Can Kurtan and Pınar Yolum. 2018. PELTE: Privacy Estimation of Images from Tags. In *Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), Stockholm, Sweden, July 10–15, 2018,* IFAAMAS, 3 pages.

1 INTRODUCTION

Online social networks (OSNs) provide personal spaces to people to share their contents, such as images, news items, and so on. Most of the time, each user prefers to share their contents with the audience that they see fit. To facilitate the sharing process, users are allowed to define the privacy setting of their content [5]. As a result, some contents are shared publicly, while some are only shared with friends. However, managing these privacy policies is difficult. Asking a user to manually set a privacy policy every time she is sharing an image will be time consuming and error prone. Further, it is possible that the user does not know which privacy settings are appropriate for a content. Automated approaches can help users to manage their privacy by predicting the privacy setting for a new image that the user wants to share.

One way to tackle this problem is to use all available images in a system—independent of who has shared the image for privacy classification [6–8]. However, it is unrealistic to assume that a single entity can access the images and the privacy policies of all users in the system. Hence, it is necessary to be able to provide estimations without requiring access to all images in the system.

Another drawback of centralized approaches is that they assume that all users share the same understanding of privacy. Since privacy is by nature subjective [4], personalized classifiers that address the preferences of a single user are needed [2, 9]. However, personalized models usually suffer from a cold start problem, initially users do not have enough data to make reliable estimations. Ideally, the system should work well even when a user has not shared many images before [9]. Further, the system should be able to adapt to changes in a user's privacy understanding and in their network. Because computation power and the knowledge of an agent is limited in a distributed system, it is unrealistic to assume that an agent can have complex learning models.

2 INFERRING PRIVACY FROM TAGS

We design a system PELTE where an agent represents a user and helps her preserve privacy when sharing images. A user can share images and can view images that are shared with her. When a user is deciding to share an image, she needs to decide with whom the image should be shared. In principle, the decision can contain various audience groups, but here for simplicity we only consider the decision of sharing publicly or privately (e.g., only with friends). Thus, a privacy setting related to an image has two values that are deny and permit, representing private and public, respectively.

PELTE aims to estimate the privacy setting of an image using its content as opposed to its metadata or other personal information of the user. Content based features of an image can be represented by its tags and automated systems can use these tags to define privacy policies [3]. Following this idea, here each image contains a set of tags that reflect its content. A tag is a keyword such as "woman" or "beach" that either identifies an object in the image or reflects a context. These tags might have been produced by the users as well as an automated tool. An agent can access the tags of the images its user has shared or has been shared with it. Each agent uses the tags to decipher what its user finds private. We propose two metrics for tags that are relevant for this purpose: *support* and *effect*.

Support value of a tag shows the number of images that have the tag. If there are many images with the same tag, we can know the privacy preference on a tag more strongly. That is, higher *support value* reveals more precise information about the user's privacy preferences on the content.

Effect value of a tag denotes the number of images with that tag that were shared publicly. Normalization of the *effect value* with

Proc. of the 17th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2018), M. Dastani, G. Sukthankar, E. André, S. Koenig (eds.), July 10–15, 2018, Stockholm, Sweden.

support value yields the ratio of public images to all (both public and private) images with the same tag. The result is between 0 and 1. If the value is smaller, images are mostly shared as private.

Above mentioned properties clearly indicate that tags with high *support value* and *effect value* that is either close to maximum or minimum value are more informative to estimate user's privacy preference for an image. If a user does not have a preference on a tag, then the tag will either not be in the overall set of tags or will have an effect value close to the average.

Each agent has its *internal tag table* to store the data of privacy settings that are collected from images that the user shares herself. The agent's *external tag table* stores the data collected from the images that the user's friends' have shared with the user. As is common with OSNs, we assume that if a user can view an image, then the user's agent can obtain the privacy settings of the image.

Each user agent starts with empty tag tables when she joins the system. Whenever a user shares a new image, the user's *internal tag table* is updated according to the privacy setting of the image. First, the agent generates tags of the images. This can be done automatically in various ways, *e.g.*, using the tool, Clarafai.com, which provides a service that takes an image and returns up to 20 tags of the image. Then, the agent updates the corresponding rows of the tags in the table. If any of the tags is not already stored in the table, it is added to the table. After the addition process, *support* and *effect* values of the tags will be updated. If the privacy setting of the image is public, then the effect values of the tags are incremented by one. Otherwise, the effect values of the tags remain the same. In both cases, the support values of the tags are incremented by one.

Confidence value is a metric to infer privacy settings of a new image from its tags. It measures the effect per support value for the tags that are associated with a given image. In doing so, it first evaluates the total effect of image tags that the agent has seen before (e.g., in the agent's overall set of tags). However, it also takes into account the image tags that the agent has not seen so far by assuming their values to be average *effect values* and *support values*. Taking into account these tags result in the metric to yield values that signal an uncertain privacy setting. This is a desired outcome because the agent has no previous experience on these tags and thus should be cautious in estimating the privacy setting.

As we mentioned before, having a confidence value that is not around the average value is more valuable to infer privacy. Therefore, the confidence value is meaningful only when it is compared with average effect value per support. Since it is possible to compare a *confidence value* of tags of an image and *average value* of all tags, the system can infer a privacy setting for the image. If the *confidence value* of the image is higher than the *average value*, the image would be considered more probable to be public. However, confidence values that are close to average could easily be misleading. To signal this to the agent, we use a threshold θ and require that the confidence is at least θ amount different than the average.

It is possible that the estimation will not reliably conclude the label as private or public. In this case, PELTE analyzes the sharing behavior of other individuals in the system. However, these are not random individuals from the network but those that the user has social ties with. For example, if two friends always share images with similar tags, this would signal that their privacy preferences are similar. Based on this intuition, in PELTE, each agent analyzes its friends' privacy settings of their shared images to judge how similar they are to each other. Agents with similar privacy preferences are favored when obtaining privacy opinions. To do that, agents calculate a trust value against their friends to compare their privacy preferences. In this context, trust is used as a measure to calculate how similar a user a's privacy preferences are with a friend, b. It compares the privacy setting of the image (as set by agent b) and the action agent a would have taken, if agent a was actually sharing the image. The similarity in an agent increases when the number of images with same privacy preferences is high.

Systems that make decisions based on historical data typically suffer from the cold start problem when the required historical data are not available. In our context, when a user has not shared images with a content, possible tags of these images cannot be found in the user's internal tag table. This can occur in two different situations: when a user's *internal tag table* does not have enough tags because she is new or she has not shared any images with that content. Therefore, estimation from internal data mechanism cannot decide to privacy setting of the image. To handle the cold part problem in the context of privacy, we estimate the privacy setting of an image from user's friends' experience on similar images via the data that can be generated by just using shared images.

In our approach, while a user agent is updating its *internal tag table*, agents of the user's friends that the image shared with, update their *external tag tables*. Now, these agents have the privacy setting of the image, which is shared with them by a friend. This time, the same update operations are performed for the *external tag table*. This process will be executed for each of the user's friends agents. Trust values are used in the update operation of external tag table. They are the multipliers of support and effect. Thus, the data of images that are shared by friends with similar privacy preferences will affect more than the data of images that are shared by others. As a result of these update processes, both the *internal tag table* and *external tag table* of the agents in an environment will be dynamically updated. More importantly, if the confidence value is around average, PELTE infers that user's preference on the image is uncertain and leave the action to the external estimation. Thus, even though a user's previous images do not have enough data itself, the system can estimate privacy settings of the image.

3 DISCUSSION

We implement PELTE on an environment where privacy settings are either private or public. However, it is possible to implement in an environment where relationship-based access control [1] is possible. Our proposed representation of the tag table can be extended with more columns that correspond to different relationship types. Thus, it is possible to define more complex privacy policies based on relationship types, which is an interesting direction that we would like to pursue in the future. Another important direction is to employ more sophisticated techniques for trust assessments. Currently, the trust assessment is done based on the similarity of users. However, since trust is multi-dimensional, an agent might trust another on certain contents but not others. Incorporating such trust reasoning into PELTE could improve its performance even further.

REFERENCES

[1] Philip WL Fong. 2011. Relationship-based access control: protection model and policy language. In *Proceedings of the first ACM conference on Data and application security and privacy*. ACM, 191–202.

[2] Berkant Kepez and Pınar Yolum. 2016. Learning privacy rules cooperatively in online social networks. In *Proceedings of the 1st International Workshop on AI for Privacy and Security*. ACM, 3.

[3] Peter Klemperer, Yuan Liang, Michelle Mazurek, Manya Sleeper, Blase Ur, Lujo Bauer, Lorrie Faith Cranor, Nitin Gupta, and Michael Reiter. 2012. Tag, you can see it!: Using tags for access control in photo sharing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 377–386.

[4] Nadin Kökciyan, Nefise Yaglikci, and Pınar Yolum. 2017. An Argumentation Approach for Resolving Privacy Disputes in Online Social Networks. *ACM Transactions on Internet Technology* 17, 3, Article 27 (June 2017), 22 pages.

[5] Nadin Kökciyan and Pınar Yolum. 2016. PriGuard: A Semantic Approach to Detect Privacy Violations in Online Social Networks. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2724–2737.

[6] Anna Squicciarini, Cornelia Caragea, and Rahul Balakavi. 2017. Toward automated online photo privacy. *ACM Transactions on the Web (TWEB)* 11, 1 (2017), 2.

[7] Jun Yu, Baopeng Zhang, Zhengzhong Kuang, Dan Lin, and Jianping Fan. 2017. iPrivacy: image privacy protection by identifying sensitive objects via deep multi-task learning. *IEEE Transactions on Information Forensics and Security* 12, 5 (2017), 1005–1016.

[8] Sergej Zerr, Stefan Siersdorfer, Jonathon Hare, and Elena Demidova. 2012. Privacy-aware image classification and search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*. ACM, 35–44.

[9] Haoti Zhong, Anna Squicciarini, David Miller, and Cornelia Caragea. 2017. A Group-Based Personalized Model for Image Privacy Classification and Labeling. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*. 3952–3958.