Privacy in Anonymous Social Media

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Lucas Kim
Spring, 2021

Technical Project Team Members
Lucas Kim
Rajiv Sarvepalli

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature __________Lucas Kim_______________________ Date _May. 8, 2021_
Lucas Kim

Approved __________Aaron Bloomfield__________________ Date _May. 8, 2021_
Aaron Bloomfield, Department of Computer Science

Privacy in Anonymous Social Media

Rajiv Sarvepalli[*]
Lucas Kim[*]
rs7uxf@virginia.edu
lk5jk@virginia.edu
University of Virginia
Charlottesville, Virginia, USA

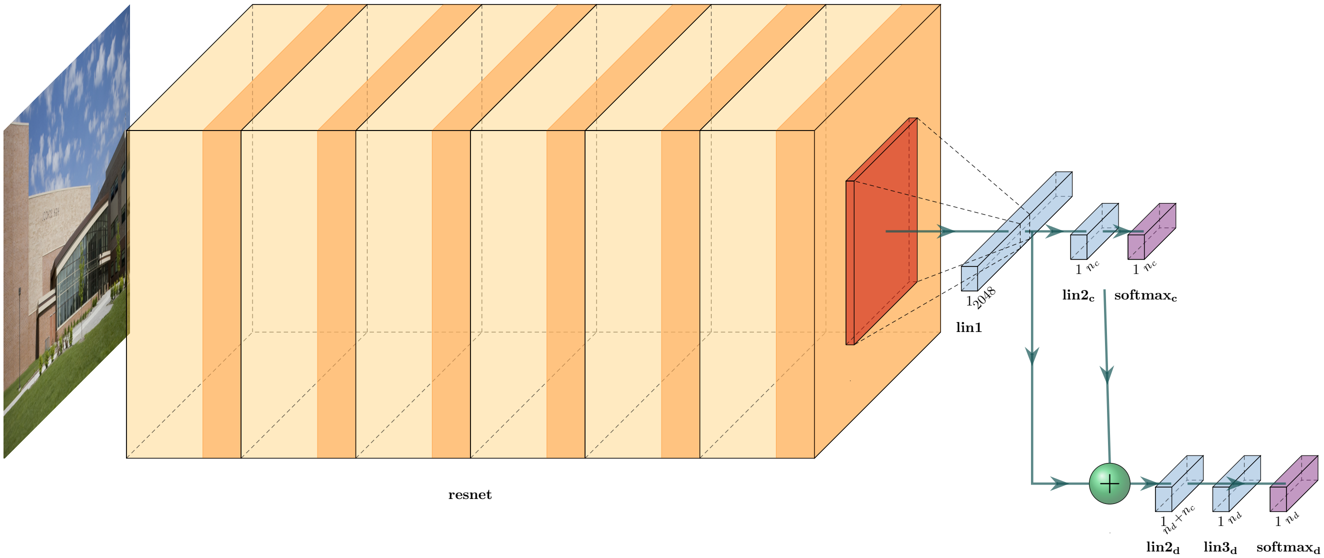

Figure 1: This is the model arhcitecure for the hierarhcal model described in 4.4. The component resnet refers to PyTorch network of a pretrained ImageNet-trained ResNeXt-101 model with cardinality 32 and a bottleneck width of 8 [7]. The n_c is the number of cities (3), and n_d is the number of districts within each city (10). The plus sign within a green ball indicates concatenation of its inputs. This figure was created using Haris Iqbal's neural network drawing tool [4].

ABSTRACT

With the rise of sites like Reddit, where individuals' personal names and information are hidden behind anonymous usernames, "anonymous" social media services have increased in popularity. However, how much information can usernames really shield? With this form of communication, to remain truly anonymous, the information from your "profile" cannot be used to determine who you are. However, with the ever-increasing data-driven world we live in, it seems the tools to link your anonymous profile with your public profiles such as Facebook, LinkedIn, and Twitter seem more and more likely. With this realization in mind, we hope to develop a tool that indicates how much personal information individuals give away in an anonymous social media post. We hope to provide a baseline for how much information anonymous social media users are giving in both a specific base and their "profile": the culmination of all their posted information.

1 INTRODUCTION

In the technological age of today, privacy becomes a more and more valuable commodity. With so many companies that live off the idea that information is money, it becomes increasingly concerning the amount of an individual's information that is public. It is public in every sense of the word, not just to a group of people, but to the whole world. Consider the constant data scandals that plague our technological world. Whether it is Facebook, Google, or governments, someone is always getting caught selling, collecting, or losing data that many consider infringes on their privacy. Therefore, as stewards of these technologies, we must develop preemptive ways of protecting the privacy of the individual in an information-based world focused on the collective. The heterogeneous nature of society, especially with respect to privacy, makes the perspective vary greatly from person to person. This study shall focus on Reddit, an anonymous social media since individuals within anonymous social media communities tend to view anonymity as some form of privacy and therefore tend to care about it. In order to understand the perspective and definitions of privacy, privacy needs to be analyzed in the context of a society.

In recent years, "anonymous" social media sources have become increasingly popular, with the rise of sites like Reddit, where true identities are typically masked with usernames. With this comes the need to scan the content you post, making sure it doesn't reveal

[*] Both authors contributed equally to this research.

anything about your identity. Our goal with this project is to manufacture a tool that allows users to scan their desired images to see whether it reveals too much personal information about themselves (represented by a privacy score). The applications of such a tool can be extended to all kinds of social media, enabling users to control and understand the amount of information they are sharing with the world. We consider a initial attempt at this concept by using the ability to determine location as the amount of personal information given away.

2 BACKGROUND

Anonymity online is popular in today's society due to the difference of interaction compared to in-person confrontation. Alongside the security and privacy that anonymity provides, all social, racial, and financial biases are hidden; these are big factors into why people crave anonymity online. Two theories perfectly explain these differences in interaction perfectly: The Equalization Phenomenon (EP) and the Social Identity Model of Deindividualization Effects (SIDE).

EP states that computer-mediated communications allow all members equal participation. Biases can't be formed as everyone is anonymous, meaning everyone appears the same online and there's no feature to be biased against. A current example would be racial hate. Someone who may act harsh towards a specific race in person wouldn't interact with that same person the same way on an anonymous forum as there's no way to distinguish the race of each other. SIDE states that anonymity advocates more for greater social identity and less personal identity. Anonymity online brings together individuals based on their mutual interests and less on their personal features. These are the reasons why Reddit is one of the most popular social media outlets and why we want to use Reddit as our focus of study for anonymous social media.

So, what is Reddit? Reddit is an online forum where all of its content is created by its anonymous users called Redditors. Redditors are distinguished solely by their username, their posts/comments, and nothing else. It's very hard to distinguish someone's reddit account to one of their public social media accounts as there's nothing to work off of based on their profile. However, the problem lies within the posts of Redditors and if information can be extracted from posts that cause Redditors to lose their anonymity. Although personal information is hidden on Reddit, personalities are not unless you never post or comment. That little bit of exposure in personal info is what can lead to a Redditor losing their anonymity and that's the problem we're trying to tackle.

3 RELATED WORK

Early work on image geotagging include [3, 6] where neural network classification and scene matching are used to predict the location of an image with no other information. All of these fail to view the problem as a hierarchical classification problem in order to identify the location. Additionally, the application of geotagging images to protect privacy is not an application that is explored.

In the context of privacy as a whole, some publications explore the technical shaping of privacy. One publication indicates the inability of private users to remain private in social media with a mix of public and private users. The public user's data of social media can be effectively used to predict the private user's behavior indicating an inability to remain truly private in most of the world's current social media [9]. One paper studies the effect of Big Data on the average user's privacy and what parts they should be concerned about. They propose a concept that will enable users to be effectively informed about the relevant privacy concerns within Big Data [5]. Whether anonymous social media is truly anonymous is analyzed by finding out if the information on anonymous social media is sufficient to track or identify users [2]. Another publication surveys the work done so far in user privacy protection and analyzes different techniques and algorithms for privacy prevention and anonymization to determine future research directions and issues [1].

Overall, the literature indicates some investigations into the problems of privacy inside social media, the failure of anonymous social media to be truly anonymous, and the different demographics perspectives on privacy. However, the current literature fails to provided tools to the user to measure the amount of information given away. While we do not intend to fully release a tool to do so, we plan to make progress on developing such a tool and design a proof-of-concept.

4 SYSTEM DESIGN

We have to design a system that determines image location from just the pixels. The obvious solution is simply to take the geocoordinates as labels, the images as input, and train some neural network that looks the solve this problem. However, this solution is too naive to work effectively. It does take full advantage of what we know the location to be. Location is simply geometrical points, and as such, they can be grouped together by distance. We can now term a classification problem in order to determine the geolocation based on groups of clustered points. However, one would expect images in Pittsburg to look different from a more modern city like images in Los Angeles. Above this level, there are images that one would expect to see in Africa, but not in America. It is impossible to be extremely accurate in this problem, but taking advantage of hierarchical classification should improve the ability to classify location. It is also to define hierarchical relationships for geocoordinates since you can simply hierarchically cluster them.

4.1 Dataset

To build an image geolocator we look to define several models that can support our task. Using a very limited image geolocation dataset, we look to define categorical location-based classifications for images. We use a dataset composed of 62,058 Google Street View images which include 3 primary cities: Pittsburgh, Orlando, and Manhattan [8]. Due to the resource and time constraints, we cannot use a dataset that includes a large number of locations, although this would be ideal. The dataset includes multiple images per location (5 per location) to cover the different angles of different views. The label distribution is not quite evenly distributed with more images being taken from Manhattan than Pittsburg and more from Pittsburg than Orlando. The lack of an even distribution makes it more challenging to train, requiring extra work to ensure that the minority classes do not have very poor accuracy. Below is a graphic displaying the label distribution for this dataset (Figure 2).

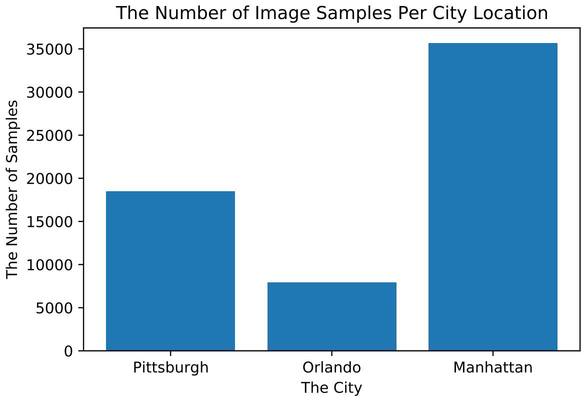

Figure 2: The number of image samples for each of the three major cities included as part of the dataset. The label distribution is not a uniform distribution, therefore introduces challenges for trained the model on a unbalanced dataset.

4.2 Labeling

The issue of labeling now comes into play when deciding what is the correct way to label an image geolocation dataset. The labels included with our dataset are simply geocoordinates (latitude and longitude). However, predicting such labels is very hard, and our model is focused on merely stating the amount of locality information given away in the model. Therefore, if it can guess with no full precision, but in the relative area, we should consider this as giving away some information. Finally, the model's confidence is what we intend to use to predict the chance of giving locality information away. Confidence makes more sense with classification tasks than with regression tasks like predicting exact geocoordinates. However, we need to figure out a means of making these classifications.

4.2.1 Clustering. The most obvious way of converting a group of geometrical coordinates into a classification problem is to cluster them into groups based on the euclidean distance between them. This intuitively makes sense, since logically, humans follow the same pattern. When we want to say a location, we do so in the form of an address. The detailed part is included first, followed by city, state, and country. The city and state are means of simply grouping close together locations. Districts and counties are other forms of doing the same concept. When humans try to guess locations based on images, I would assume most to try to guess broader locations like countries first, then try to narrow it down. Considering this is a model that allows humans to efficiently communicate and guess locations, why not use the same idea for our dataset labeling?

There are several methods of organizing the clustering that can be done to find groupings of geocoordinates for our images. However, considering the model that people often use, it makes sense to do hierarchical clustering. We want to create clusters, and then create clusters inside each cluster to provide a multi-label system with increased granularity than just singular clusters. Otherwise, with singular clusters, the problem becomes too simple or illogical (depending on the number of clusters). Therefore, we define our K-means clustering to cluster twice with increasing granularity. We use a number of clusters that seem reasonable based on the provided data. The first clustering level uses 3 clusters because there are 3 widely spaced-apart cities. Then, for each of those 3 clusters, we define a set of 10 clusters which we term districts. Note that these districts are not actually districts within the city, but rather 10 K-means clusters inside the city. From now on these 10 clusters within each of the larger 3 clusters will be referred to as districts. This gives us two "classes" to predict on, both the city and districts. We will discuss more in the model's section how to encode this class structure into a meaningful model to extract as much information as possible from the data and the created labels. Below is an image that displays a graphic for how the clustering was done to create groupings in both the cities and the districts for one of the cities (Figure 3).

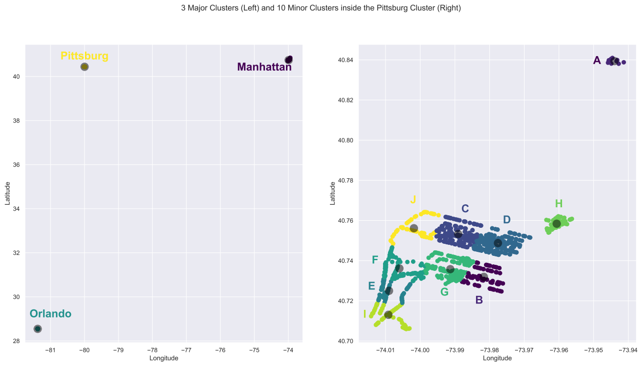

Figure 3: On the left is a clustering of the locations into initial categories, and on the right is clustering of one of those initial categories. The shown points are 1,000 data samples in each depiction. The left side displays the city clusters and the right side is the district clusters inside of the Pittsburg city cluster.

4.3 Class Connections

The new class labels we have constructed share information between them. The city label inherently limits the number of potential district labels to 10. This pattern is essential to think about when designing a model that can use this kind of information. In the same manner, in which a human thinks broadly first, and narrows down a location, an ideal model would guess the city location first and then feed it forward to determine the district classification. Therefore, we can view the class as a hierarchy itself with information being coded downstream from the parents to the child. Our simple combination of cities and districts is a simple enough hierarchy, but when creating hierarchies for a more complicated set of locations, perhaps there should be many levels to the hierarchy. We intend to construct a model that will suit this kind of scenario, despite our hierarchy only having the city as a parent set of classes and district as a child.

The multi-set of options with differing numbers of cities or districts which could become more complex with more locations available introduces some challenges for creating a model architecture.

The model section will dictate some of these challenges along with the different means of handling this structure.

4.4 Models

Several different ways are considered for creating categorical labels from geographical coordinates for each image for classification. Two main ways can be considered:

(1) Splitting based on hierarchical clustering
(2) Splitting based on geographical separations (such as cities)

The first method is an obvious way of splitting data, but to give our model more understanding of what composes locality information in an image, the data can be clustered in a hierarchy. This will increase the ability to understand privacy information since we are considering privacy with respect to locality information. As explained before, The first clustering will be cities and a secondary grouping will be split between those cities (called district)s. In Figure 3, we demonstrated how our K-Means model was hierarchically clustered cities into reasonable districts for the classification problem. Our major objective is to increase the accuracy of classifying districts rather than classifying cities, because it is a more complex problem, and will reinforce our model's ability to understand geographical location. Additionally, it will force the model to pick up on more subtle details, and thus create a better prototype.

The score for privacy will be 1− the district confidence or, in other words, the complement of the district confidence. This is due to the fact that we assume when the model is less certain, the amount of locality information given away is less. Essentially, the assumption is that the confidence of the model and the about of locality information are linearly correlated. The calculation for the privacy score will be assumed to be 1 if the city confidence is below some threshold. The intuition here is that if the model cannot confidently predict the city (a relatively simple task), then it could know the location of the image accurately. Therefore, the image is quite unknown and will have a privacy score of 1. The formula for the privacy score is displayed below with respect to an image. γ is the city confidence threshold, α is the city confidence, and β is the district confidence.

$$\text{Privacy}_{\text{image}}(\gamma, \beta, \alpha) = \begin{cases} 1 & \text{if } \alpha < \gamma \\ 1 - \beta & \text{if } \alpha \geq \gamma \end{cases}$$

Again, γ is the cutoff value chosen for the model's confidence score for cities.

The three different architectures are tested to illustrate the different methods of considering this problem of classifying hierarchical data. Additionally, it enables us to compare the different methodologies. Initially, we assume the problem to take form as multi-class with multiple classes that could be true. In this case, the model includes a final sigmoid layer and has 33 outputs (3 cities + 3*10 districts). This model outputs into one-hot vector encoding. However, this model does not make sense, because it assumes that multiple city classifications are possible. Despite this, we simply include it as a point of comparison and as a naive solution to the problem. Second, we use PyTorch to simply have our model output two different outputs: a district and city prediction. This model is trained using the summed losses of each classification when compared to the relevant label. Finally, we use the hierarchical where the outputs of the parent classes are feedback to be inputs for a child class. This model is defined using a class hierarchy and created through simple-hierarchy, a python library. All three of these models use a back-end model of a pre-trained ImageNet-trained ResNeXt-101 model with cardinality 32 and a bottleneck width of 8 [7]. The model architecture of the hierarchical model is described in Figure 1. The green circles with a plus indicate concatenation of the inputs to the circle. We predict that the hierarchical model will perform the best given its architecture is connected to the relationship between the hierarchical class structure. The Python library simple-hierarchy was created as a part of this project for making it simple to create hierarchical models to replicate models that are similar to the one used for this specific instance. The library has a variety of parameters enabling a wider set of hierarchical models to be created for all kinds of tasks, but that follows the idea of the class structure following some hierarchy. The library includes a GitHub repository and documentation as well. The documentation provides several examples and should illustrate what the library is capable of along with how to use it.

Despite the fact that we are mainly concerned with the network's ability to recognize locality information from the images, we still use location accuracy as our performance metric. Although this is not entirely the same thing, we predict that a model that can more accurately predict location can more accurately recognize features that give away location information. The hierarchical classifications could benefit from a more complex model that better grasp the nature of the hierarchy and/or a dataset that helps define essential features for determining location. At the end of the day, we are concerned with saying how much locality information is given away rather than accurately predicting the location, but the former is much a harder thing to label.

However, another important model architecture that could be used is using two separate models for city and district predictions and feeding the city output into the district model prediction. This perhaps would achieve the highest accuracy, but it is not practical since it requires a new model for each set of classes. Therefore, we omit this model architecture due to computing and time constraints, despite it being potentially a high-accuracy architecture.

5 PROCEDURE

This is a hard problem, and as such, due to computing, time, and data constraints, as stated before, we reduce the problem the three major cities Manhattan, Pittsburg, and Orlando. The dataset supporting this project includes these three cities.

5.1 Experimental Setup

We prepare the data-set with standard normalization and preprocessing, and then test three different models. The first model is simply considering the problem as a multi-class problem where multiple classes can be true. The second model uses PyTorch to predict two different outputs and trains on the combined loss function. The third model is our hierarchical model creating using the library described above. Each model was given 10 epochs to train on the training dataset of (~ 50,000 images(training data was 80% of the total data).

Method, Class	Precision	Recall	F1-Score
Multi-Label, District	0.256	0.220	0.201
2 Outputs, District	0.712	0.735	0.716
Hierarchical, District	**0.740**	**0.755**	**0.743**
Multi-Label, City	0.848	0.823	0.834
2 Outputs, City	0.991	0.982	0.986
Hierarchical, City	**0.993**	**0.991**	**0.992**

Table 1: Performance metrics from the three different models described in 4.4. Each one was trained for 10 epochs with learning rate of 0.0001 **using 80% of the total dataset as training (~ 50,000 images). For each model arch, the first section includes the district classification performance and the second section includes city classification performance. The hierarchical model outperformed the other models for all metrics measured in all categories (albeit by a small margin).**

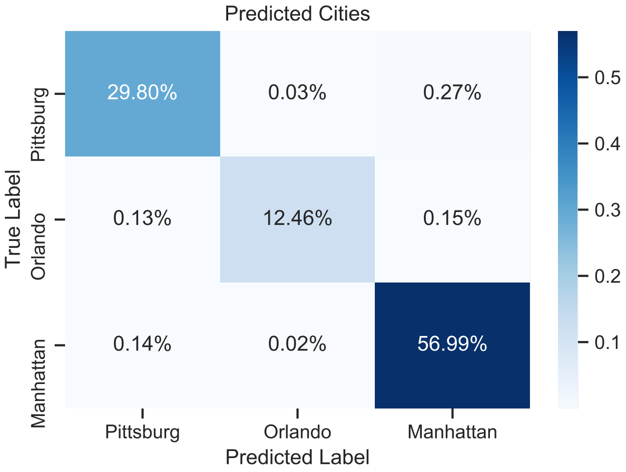

Figure 4: The confusion matrix for the hierarchical architecture for the city predictions. The annotated percents are the percentage of the total validation dataset (~ 12,000 images in the validation dataset).

6 RESULTS

Using the three models described in the model section, the performance is reported in 1 on the validation dataset. The performance metrics reported are precision, recall, and f1-score (average of recall and precision) for each class category. The hierarchical model performs best on both cities and districts classifications. These were the expected results especially for districts since the hierarchical model feeds forward the city prediction to make the district prediction. However, the gap between the 2 output models and the hierarchical model is quite small. The gap might increase when the separate portions of a network are increased (making it closer to two networks than one), but due to time and computing constraints, our hierarchical network was limited to being quite small (as viewed by the model architecture in Figure 1).

The confusion matrix for the hierarchical model architecture is shown in Figure 4 for only city predictions. The model can accurately predict cities, but this is expected given the dataset has a small subset of all the cities in the world. Classifying the districts correctly is a more challenging task and therefore better indicates how well the model is performing. The accuracies displayed in Table 1 show a reasonable high district accuracy, which makes it more likely that the model understands what components of the image are revealing with respect to location.

7 CONCLUSIONS

Based on our three models, we can determine that the Hierarchical model is able to predict city origins of posts with the highest confidence level out of the three models. In conclusion, using our Hierarchical Model, the ability to create a tool that allows us to accurately alert the user on how publicly exposing their posts can be with a high confidence rate is feasible and will help our case on increasing anonymity online, specifically relating to our research interest (Reddit).

8 FUTURE WORK

The models created here indicate a proof-of-concept of building a system that indicates a privacy score providing users with a utility to understand how much information they are giving away. There are several obvious steps forward: including more cities for the image locator models and adding text geolocation. More importantly, the future steps should really be trying to build a system that connects anonymous accounts with public profiles. However, it is very hard to create or find a dataset for this task, but ideally, we would communicate to users that we have identified you based on these images/words, and here is what you can do to remain anonymous. That would be the final design goal of continued work.

REFERENCES

[1] Ghazaleh Beigi and Huan Liu. 2018. Privacy in social media: Identification, mitigation and applications. *arXiv preprint arXiv:1808.02191* (2018).

[2] Vasileios Chatzistefanou and Konstantinos Limniotis. 2019. Anonymity in social networks: the case of anonymous social media. *International Journal of Electronic Governance* 11, 3-4 (2019), 361–385.

[3] James Hays and Alexi Efros. [n.d.]. Estimating Geographical Information From a Single Image. ([n. d.]). http://graphics.cs.cmu.edu/projects/im2gps/im2gps.pdf

[4] Haris Iqbal. 2018. PlotNeuralNet v1.0.0. https://github.com/HarisIqbal88/PlotNeuralNet.

[5] Matthew Smith, Christian Szongott, Benjamin Henne, and Gabriele Von Voigt. 2012. Big data privacy issues in public social media. In *2012 6th IEEE international conference on digital ecosystems and technologies (DEST)*. IEEE, 1–6.

[6] Nam N. Vo, Nathan Jacobs, and James Hays. 2017. Revisiting IM2GPS in the Deep Learning Era. *CoRR* abs/1705.04838 (2017). arXiv:1705.04838 http://arxiv.org/abs/1705.04838

[7] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He. 2017. Aggregated Residual Transformations for Deep Neural Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5987–5995. https://doi.org/10.1109/CVPR.2017.634

[8] A.R. Zamir and M. Shah. 2014. Image Geo-localization Based on Multiple Nearest Neighbor Feature Matching using Generalized Graphs. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* PP, 99 (2014), 1–1. https://doi.org/10.1109/TPAMI.2014.2299799

[9] Elena Zheleva and Lise Getoor. 2009. 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*. 531–540.