To Join or Not to Join: The Illusion of Privacy in Social Networks with Mixed Public and Private User Profiles

Elena Zheleva
Department of Computer Science
University of Maryland, College Park
elena@cs.umd.edu

Lise Getoor
Department of Computer Science
University of Maryland, College Park
getoor@cs.umd.edu

ABSTRACT

In order to address privacy concerns, many social media websites allow users to hide their personal profiles from the public. In this work, we show how an adversary can exploit an online social network with a mixture of public and private user profiles to predict the private attributes of users. We map this problem to a relational classification problem and we propose practical models that use friendship and group membership information (which is often *not* hidden) to infer sensitive attributes. The key novel idea is that in addition to friendship links, groups can be carriers of significant information. We show that on several well-known social media sites, we can easily and accurately recover the information of private-profile users. To the best of our knowledge, this is the first work that uses link-based and group-based classification to study privacy implications in social networks with mixed public and private user profiles.

Categories and Subject Descriptors

H.2.8 [**Information Systems**]: Data Mining

General Terms

Algorithms, Experimentation

Keywords

privacy, social networks, groups, attribute inference

1. INTRODUCTION

In order to address users' privacy concerns, a number of social media and social network websites, such as Facebook, Orkut and Flickr, allow their participants to set the privacy level of their online profiles and to disclose either some or none of the attributes in their profiles. While some users make use of these features, others are more open to sharing personal information. Some people feel comfortable displaying personal attributes such as age, political affiliation or location, while others do not. In addition, most social-media users utilize the social networking services provided by forming friendship links and affiliating with groups of interest. While a person's profile may remain private, the friendship links and group affiliations are often visible to the public. Unfortunately, these friendships and affiliations leak information; in fact, as we will show, they can leak a surpisingly large amount of information.

Copyright is held by the International World Wide Web Conference Committee (IW3C2). Distribution of these papers is limited to classroom use, and personal use by others.
WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

The problem we consider is *sensitive attribute inference* in social networks: inferring the private information of users given a social network in which some profiles and all links and group memberships are public (this is a commonly occurring scenario in existing social media sites). We define the problem formally in Section 4. We believe our work is the first one to look at this problem, and to map it to a relational classification problem in network data with groups.

Here, we propose eight privacy attacks for sensitive attribute inference. The attacks use different classifiers and features, and show different ways in which an adversary can utilize links and groups in predicting private information. We evaluate our proposed models using sample datasets from four well-known social media websites: Flickr, Facebook, Dogster and BibSonomy. All of these websites allow their users to form friendships and participate in groups, and our results show that attacks using the group information achieve significantly better accuracy than the models that ignore it. This suggests that group memberships have a strong potential for leaking information, and if they are public, users' privacy in social networks is illusionary at best.

Our contributions include the following:

- We identify a number of novel privacy attacks in social networks with a mixture of public and private profiles.
- We propose that in addition to friendship links, group affiliations can be carriers of significant information.
- We show how to reduce the large number of potential groups in order to improve the attribute accuracy.
- We evaluate our attacks on challenging classification tasks in four social media datasets.
- We illustrate the privacy implications of publicly affiliating with groups in social networks and discuss how our study affects anonymization of social networks.
- We show how surprisingly easy it is to infer private information from group membership data.

We motivate the problem in the next section. Then, we describe the data model in Section 3. Section 4 presents the privacy attacks, and Section 5 provides experimental results using these attacks. Section 6 presents related work, and Section 7 discusses the broader implications of our results.

2. MOTIVATION

Disclosing private information means violating the rights of people to control who can access their private informa-

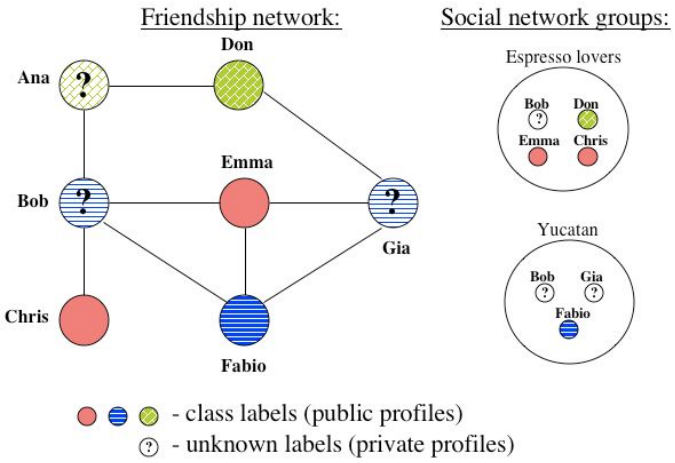

Figure 1: Toy instance of the data model.

tion. In order to prevent private information leakage, it is important to be aware of the ways in which an adversary can attack a social network to learn users' private attributes. Studies on the challenges of preserving the privacy of individuals in social networks have emerged only in the last few years, and they have concentrated on inferring the identity of nodes based on structural properties such as node degree. In contrast, we are interested in inferring sensitive attribute of nodes using approaches developed for relational learning, another active area of research in the last few years.

The novelty of our work is that we study the implications of mixing private and public profiles in a social network. For example, in Facebook many users choose to set their profiles to private, so that no one but their friends can see their profile details. Yet, fewer people hide their friendship links and even if they do, their friendship links can be found through the backlinks from their public-profile friends. Similarly for group participation information – even if a user makes her profile private, her participation in a public group is shown on the group's membership list. Currently, neither Facebook nor Flickr allow users to hide their group memberships from public groups. Both commercial and governmental entities may employ privacy attacks for targeted marketing, health care screening or political monitoring – just to mention a few. Therefore, social media website providers need to protect their users against undesired eavesdropping and inform them of the possible privacy breaches and providing them with the means to be in full control of their private data.

Our work is also complimentary to work on data anonymization, in which the goal is to perturb data in such a way that the privacy of individuals is preserved. Our goal is not to release anonymized data but to illustrate how social network data can be exploited to predict hidden information: an essential knowledge in the anonymization process.

We identify a new type of privacy breach in relational data, *group membership disclosure*: whether a person belongs to a group relevant to the classification of a sensitive attribute. We conjecture that group membership disclosure can lead to attribute disclosure. Thus, hiding group memberships is a key to preserving the privacy of individuals.

3. DATA MODEL

We represent a social network as a graph $G = (V, E, H)$, where V is a set of n nodes of the same type, E is a set of edges (the friendship links), and H is a set of groups that nodes can belong to. $e_{i,j} \in E$ represents a directed link from node v_i to node v_j. Our model handles undirected links by representing them as pairs of directed links. We describe a group as a hyper-edge $h \in H$ among all the nodes who belong to that group; $h.U$ denotes the set of users who are connected through hyper-edge h and $v.H$ denotes the groups that node v belongs to. Similarly, $v.F$ is the set of nodes that v has connected to: $v_i.F = \{v_j | \exists e_{i,j} \in E\}$. A group can also have a set of properties $h.T$.

We assume that each node v has a sensitive attribute $v.a$ which is either observed or hidden in the data. A *sensitive attribute* is a personal attribute, such as age, political affiliation or location, which some users in the social network are willing to disclose publicly. A sensitive attribute value can take on one of a set of possible values $\{a_1 ... a_m\}$. A *user profile* has a unique id with which the user forms links and participates in groups. Each profile is associated with a sensitive attribute, either observed or hidden. A *private profile* is one for which the sensitive attribute value is unknown, and a *public profile* is the opposite: a profile with an observed sensitive attribute value. We refer to the set of nodes with private profiles as the *sensitive set* of nodes V_s, and to the rest as the *observed set* V_o. The adversary's goal is to predict $V_s.A$, the sensitive attributes of the private profiles.

Here, we study the case where nodes have no other attributes beyond the sensitive attribute. Thus, to make inferences about the sensitive attribute, we need to use some form of relational classifier. While additional attribute information can be helpful and many relational classifiers can make use of it, in our setting this is not possible because all of the private-profile attributes are likely to be hidden.

As a running example, we consider the social network presented in Figure 1. It describes a collection of individuals (Ana, Bob, Chris, Don, Emma, Fabio, and Gia), along with their friendship links and their groups of interest. Chris, Don, Emma and Fabio are displaying their attribute values publicly, while Ana, Bob and Gia are keeping theirs private. Emma and Chris have the same sensitive attribute value (marked solid), Bob, Gia and Fabio share the same attribute value (marked with stripes), and Ana and Don have a third value (marked with a brick pattern). Users are linked by a friendship link, and in this example they are reciprocal. There are two groups that users can participate in: the "Espresso lovers" group and the "Yucatan" group. While affiliating with some groups may be related to the sensitive attribute, affiliating with others is not. For example, if the sensitive attribute is a person's country of origin, the "Yucatan" group may be relevant. Thus, this group can leak information about sensitive attributes, although the manner in which it is leaked is not necessarily straightforward.

4. SENSITIVE-ATTRIBUTE INFERENCE MODELS

The attributes of users who are connected in social networks are often correlated. At the same time, online communities allow very diverse people to connect to each other and form relationships that transcend gender, religion, origin and other boundaries. As this happens, it becomes harder to utilize the complex interactions in online social networks for predicting user attributes.

Attribute disclosure occurs when an adversary is able to

infer the sensitive attribute of a real-world entity accurately. The sensitive attribute value of an individual can be modeled as a random variable. This random variable's distribution can depend on the overall network's attribute distribution, the friendship network's attribute distribution and/or the attribute distribution of each group the user joins.

The problem of *sensitive attribute inference* is to infer the hidden sensitive values, $V_s.A$, conditioned on the observed sensitive values, links and group membership in graph G. We assume that the adversary can apply a probabilistic model M for predicting the hidden sensitive attribute values, and he can combine the given graph information in various ways as we discuss next. The prediction of each model is:

$$v_s.\hat{a}_M = \operatorname*{argmax}_{a_i} P_M(v_s.a = a_i; G).$$

where $P_M(v_s.a = a_i; G)$ is the probability that the sensitive attribute value of node $v_s \in V_s$ is a_i according to model M and the observed part of graph G.

We assume that the overall distribution of the sensitive attribute is either known or it can be found using the public profiles. An attack using this distribution is a *baseline attack*. A *successful attack* is one which, given extra knowledge, e.g., friendship links or group affiliations, has a significantly higher accuracy than the baseline attack. The extra knowledge *compromises* the privacy of users if there is an attack which uses it and is successful.

4.1 Attacks without links and groups

In the absence of relationship and group information, the only available information is the overall marginal distribution for the sensitive attribute in the public profiles. So, the simplest model is to use this as the basis for predicting the sensitive attributes of the private profiles. More precisely, according to this model, BASIC, the probability of a sensitive attribute value can be estimated as the fraction of observed users who have that sensitive attribute value:

$$P_{BASIC}(v_s.a = a_i; G) = P(v_s.a = a_i | V_o.A) = \frac{|V_o.a_i|}{|V_o|},$$

where $|V_o.a_i|$ is the number of public profiles with sensitive attribute value a_i and $|V_o|$ is the total number of public profiles. The adversary using model BASIC picks the most probable attribute value which in this case is the overall mode of the multinomial attribute distribution. In our toy example, the most common observed sensitive attribute is the value that Chris and Emma share. Therefore, the adversary would predict that Ana, Bob and Gia have the same attribute value as well. An obvious problem with this approach is that if there is a sensitive attribute value that is predominant in the observed data, it will be predicted for all users with private profiles. Nevertheless, this attack is always at least as good as a random guess, and we use it as a simple baseline. Next, we look at using friendship information for inferring the attribute value.

4.2 Privacy attacks using links

Link-based privacy attacks take advantage of *autocorrelation*, the property that the attribute values of linked objects are correlated. An example of autocorrelation is that people who are friends often share common characteristics (as in the proverb "Tell me who your friends are, and I'll tell you who you are"). Figure 2(a) shows a graphical representation

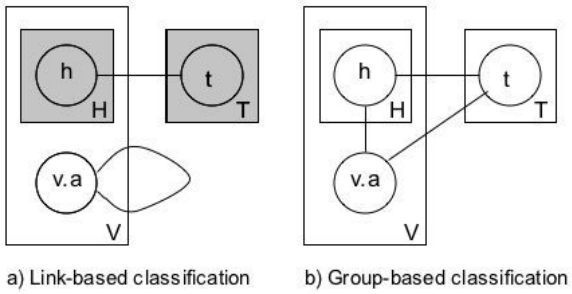

Figure 2: Graphical representation of the models. Grayed areas correspond to variables that are ignored in the model.

of the link-based classification model. There is a random variable associated with each sensitive attribute $v.a$, and the sensitive attributes of linked nodes are correlated. The greying of the other two types of random variables means that the group information is not used in this model.

4.2.1 Friend-aggregate model (AGG)

The nodes and their links produce a graph structure in which one can identify circles of close friends. For example, the circle of Bob's friends is the set of users that he has links to: $Bob.F = \{Ana, Chris, Emma, Fabio\}$. The friend-aggregate model AGG looks at the sensitive attribute distribution amongst the friends of the person under question. According to this model, the probability of the sensitive attribute value can be estimated by:

$$P_{AGG}(v_s.a = a_i; G) = P(v_s.a = a_i | V_o.A, E) = \frac{|V'_o.a_i|}{|V'_o|}$$

where $V'_o = \{v_o \in V_o | \exists (v_s, v_o) \in E\}$ and $V'_o.a_i = \{v_o \in V'_o | v_o.a = a_i\}$.

Again, the adversary using this model picks the most probable attribute value (i.e., the mode of the friends' attribute distribution). In our toy example (Figure 1), Bob would pick the same value as Emma and Chris, Ana the same label as Don, and Gia will be undecided between Don's, Emma's and Fabio's label. One problem with this method is the one when person's friends are very diverse, as in Gia's case, it will be difficult to make a prediction.

4.2.2 Collective classification model (CC)

Collective classification also takes advantage of autocorrelation between linked objects. Unlike more traditional methods, in which each instance is classified independently of the rest, collective classification aims at learning and inferring class labels of linked objects together. In our setting, it makes use of not only the public profiles but also the inferred values for connected private profiles. Collective classification has been an active area of research in the last decade (see Sen et al. [21] for a survey). Some of the approximate inference algorithms proposed include iterative classification (ICA), Gibbs sampling, loopy belief propagation and mean-field relaxation labeling.

For our experiments, we have chosen to use ICA because it is simple, fast and has been shown to perform well on a number of problems [21]. In our setting, ICA first assigns a label to each private profile based on the labels of the friends with public profiles, then it iteratively re-assigns labels considering the labels of both public and private-profile friends.

The assignment is based on a local classifier which takes the friends' class labels as features. For example, a simple classifier could assign a label based on the majority of the friends labels. A more sophisticated classifier can be trained using the counts of friends' labels.

4.2.3 Flat-link model (LINK)

Another approach to dealing with links is to "flatten" the data by considering the adjacency matrix of the graph. In this model, each row in the matrix is a user instance. In other words, each user has a list of binary features of the size of the network, and each feature has a value of 1 if the user is friends with the person who corresponds to this feature, and 0 otherwise. The user instance also has a class label which is known if the user's profile is public, and unknown if it is private. The instances with public profiles are the training data which can be fed to any traditional classifier, such as Naïve Bayes, logistic regression or SVM. The learned model can then be applied to predict the private profile labels.

4.2.4 Blockmodeling attack (BLOCK)

The next category of link-based methods we explored are approaches based on blockmodeling [24, 2]. The basic idea behind *stochastic blockmodeling* is that users form natural clusters or blocks, and their interactions can be explained by the blocks they belong to. In particular, the link probability between two users is the same as the link probability between their corresponding blocks. If sensitive attribute values separate users into blocks, then based on the observed interactions of a private-profile user with public-profile users, one can predict the most likely block the user belongs to and thus discover the attribute value. Let block B_i denote the set of public profiles that have attribute value a_i, and $\lambda_{i,j}$ the probability that a link exists between users in block B_i and users in block B_j. Thus, λ_i is the vector of all link probabilities between block B_i and each block $B_1, ..., B_m$. Similarly, let the probability of a link between a single user v and a block B_j be $\lambda(v)_j$ with $\lambda(v)$ being the vector of link probabilities between v and each block. To find the probability that a private-profile user belongs to a particular block, the model looks at the maximum similarity between the interaction patterns (link probability to each block) of the node in question and the overall interactions between blocks. After finding the most likely block, the sensitive attribute value is predicted. The probability of an attribute value using the blockmodeling attack, BLOCK, is estimated by:

$$P_{BLOCK}(v_s.a_i; G) = P(v_s.a_i | V_o.A, E, \lambda) = \frac{1}{Z} sim(\lambda_i, \lambda(v))$$

where $sim()$ can be any vector similarity function and Z is a normalization factor. We compute maximum similarity using the minimum L2 norm. This model is similar to the class-distribution relational-neighbour classifier described in [17] when the weight of each directed edge is inversely proportional to the size of the class of the receiving node.

4.3 Privacy attacks using groups

In addition to link or friendship information, social networks offer a very rich structure through the group memberships of users. All individuals in a group are bound together by some observed or hidden interest(s) that they share, and individuals often belong to more than one group. Thus, groups offer a broad perspective on a person, and it may be possible to use them for sensitive attribute inference. If a user belongs to only one group (as it is Gia's case in the toy example), then it is straightforward to infer a label using an aggregate, e.g., the mode, of her groupmates' labels, similar to the friend-aggregate model. This problem becomes more complex when there are multiple groups that a user belongs to, and their distributions suggest different values for the sensitive attribute. We propose two models for utilizing the groups in predicting the sensitive attribute – a model which assumes that all groupmates are friends and one which takes groups as classifier features.

4.3.1 Groupmate-link model (CLIQUE)

One can think of groupmates as friends to whom users are implicitly linked. In this model, we assume that each group is a clique of friends, thus creating a friendship link between users who belong to at least one group together. This data representation allows us to apply any of the link-based models that we have already described. The advantage of this model is that it simplifies the problem to a link-based classification problem, which has been studied more thoroughly. One of the disadvantages is that it doesn't account for the strength of the relationship between two people, e.g. number of common groups.

4.3.2 Group-based classification model (GROUP)

Another approach to dealing with groups is to consider each group as a feature in a classifier. While some groups may be useful in inferring the sensitive attribute, a problem in many of the datasets that we encountered was that users were members of a very large number of groups, so identifying which groups are likely to be predictive is a key. Ideally, we would like to discard group memberships irrelevant to the classification task. For example, the group "Yucatan" may be relevant for finding where a person is from, but "Espresso lovers" may not be.

To select the relevant groups, one can apply standard feature selection criteria [14]. If there are N groups, the number of candidate group subsets is 2^N, and finding an optimal feature subset is intractable. Similar to pruning words in document classification, one can prune groups based on their properties and evaluate their predictive accuracy. Example group properties include density, size and homogeneity. Smaller groups may be more predictive than large groups, and groups with high homogeneity may be more predictive of the class value. For example, if the classification task is to predict the country that people are from, a cultural group in which 90% of the people are from the same country is more likely to be predictive of the country class label. One way to measure group homogeneity is by computing the entropy of the group: $Entropy(h) = -\sum_{i=1}^{m} p(a_i) \log_2 p(a_i)$ where m is the number of possible node class values and $p(a_i)$ is the fraction of observed members that have class value a_i: $p(a_i) = \frac{|h.V.a_i|}{|h.V|}$.

For example, the group "Yucatan" has an entropy of 0 because only one attribute value is represented there, therefore its homogeneity is very high. We also consider the confidence in the computed group entropy. One way to measure this is through the percent of public profiles in the group.

The group-based classification approach contains three main steps as Algorithm 1 shows. In the first step, the algorithm performs feature selection: it selects the groups that

are relevant to the node classification task. This can either be done automatically or by a domain expert. Ideally, when the number of groups is high, the feature selection should be automated. For example, the function $isRelevant(h)$ can return *true* if the entropy of group h is low. In the second step, the algorithm learns a global function f, e.g., trains a classifier, that takes the relevant groups of a node as features and returns the sensitive attribute value. This step uses only the nodes from the observed set whose sensitive attributes are known. Each node v is represented as a binary vector where each dimension corresponds to a unique group: $\{groupId : isMember\}$, $v.a$. Only memberships to relevant groups are considered and $v.a$ is the class coming from a multinomial distribution which denotes the sensitive-attribute value. In the third step, the classifier returns the predicted sensitive attribute for each private profile. Figure 2(b) shows a graphical representation of the group-based classification model. It shows that there is a dependence between the nodes' sensitive attributes $V.A$, the group memberships H and the group attributes T.

Algorithm 1 Group-based classification model

```
Set of relevant groups H_relevant = ∅
for each group h ∈ H do
    if isRelevant(h) then
        H_relevant = H_relevant ∪ {h}
    end if
end for
trainClassifier(f, V_o, H_relevant)
for each sensitive node v ∈ V_s do
    v.â = f(v.H_relevant)
end for
```

4.4 Privacy attacks using links and groups

It is possible to construct a method which uses both links and groups to predict the sensitive attributes of users. We use a simple method which combines the flat-link and the group-based classification models into one: LINK-GROUP. This model uses all links and groups as features, thus utilizing the full power of available information. Like LINK and GROUP, LINK-GROUP can use any traditional classifier.

5. EXPERIMENTS

We evaluated the effectiveness of each of the proposed models for inferring sensitive attributes in social networks.

5.1 Data description

For our evaluation, we studied four diverse online communities: the photo-sharing website Flickr, the social network Facebook, Dogster, an online social network for dogs, and the social bookmarking system BibSonomy[1]. Table 1 shows properties of the datasets, including the sensitive attributes.

Flickr is a photo-sharing community in which users can display photographs, create directed friendship links and participate in groups of common interest. Users have the choice of providing personal information on their profiles, such as gender, marital status and location. We collected a snowball sample of 14,451 users from it. To resolve their locations (which users enter manually, as opposed to choosing them from a list), we used a two-step process. First, we used Google Maps API[2] to find the latitude and longitude of each location. Then, we mapped the latitude and longitude back to a country location using the reverse-geocoding capabilities of GeoNames[3]. We discarded the profiles with no resolved country location (34%), and ones that belonged to a country with less than 10 representatives. The resulting sample contained 9,179 users from 55 countries. There were 47,754 groups with at least 2 members in the sample.

Facebook is a social network which allows users to communicate with each other, to form undirected friendship links and participate in groups and events. We used a part of the Facebook network, available for research purposes [10]. It contains all 1,598 profiles of first-year students in a small college. The dataset does not contain group information but it contains the favorite books, music and movies of the users, and we considered them to be the groups that unify people. 1,225 of the users share at least one group with another person, and 1,576 users have friendship links. All profiles have gender and 965 have self-declared political views. We use six labels of political views - *very liberal or liberal* (545 profiles), *moderate* (210), *conservative or very conservative* (114), *libertarian* (29), *apathetic* (18), and *other* (49).

Dogster is a website where dog owners can create profiles describing their dogs, as well as participate in group memberships. Members maintain links to friends and family. From a random sample of 10,000 Dogster profiles, we removed the ones that do not participate in any groups. The remaining 2,632 dogs participate in 1,042 groups with at least two members each. Dogs have breeds, and each breed belongs to a broader type set. In our dataset, there were mostly *toy* dogs (749). The other breed categories were *working* (268), *herding* (202), *terrier* (232), *sporting* (308), *non-sporting* (225), *hound* (152) and *mixed dogs* (506).

The fourth dataset contains publicly available data from the social bookmarking website BibSonomy[4], in which users can tag bookmarks and publications. Although BibSonomy allows users to form friendships and join groups of interest, the dataset did not contain this information. Therefore, we consider each tag placed by a person to be a group to which a user belongs. There are no links between users other than the group affiliations. There are 31,715 users with at least one tag, 98.7% of which posted the same tag with at least one other user. The sensitive attribute is the binary attribute of whether someone is a spammer or not.

5.2 Experimental setup

We ran experiments for each of the presented attack models: 1) the baseline model, an attack in the absence of link and group information (BASIC), 2) the friend-aggregate attack (AGG), 3) the collective classification attack (CC), 4) the flat-link attack (LINK) and 5) the blockmodeling attack (BLOCK), 6) the groupmate-link attack (CLIQUE), 7) the group-based classification attack (GROUP) and 8) the attack which uses both links and groups (LINK-GROUP). For the GROUP model, we present results on both the simpler version which considers all groups and the method in which relevant groups are selected. For the BLOCK model, we present leave-one-out experiments assuming that complete information is given in the network in order to predict the sensitive-attribute of a user. For the AGG, CC, LINK,

[1]At `http://www.flickr.com`, `http://www.facebook.com`, `http://www.dogster.com`, `http://www.bibsonomy.org/`

[2]At `http://code.google.com/apis/`.

[3]At `http://www.geonames.org/export/`.

[4]At `http://www.kde.cs.uni-kassel.de/ws/rsdc08/`.