

# Structural Similarity in Social Networks with Identity Separation

Zoé Rimay

The research of networks has become very popular in the last few decades, mainly because it can be used in almost any field of science. Nowadays, actively participating in online networks – like social networking websites, blogs or professional networks – is very widespread, as well as their scientific investigation. The structural similarity of the participants of such networks can be very interesting from several points of view, whether it is about the re-identification (or de-anonymization) of the anonymized users of these networks or simply just about a referral system. Moreover, the traditional social networks have a lot of similarity measures, used in other areas of science (e.g. biology, economics, sociology), as well. These measures are the main topic of my thesis.

De-anonymization of the dataset can be prevented with several techniques, like adding some noise to the profiles or structures, merging or transforming some parts of the networks. However, these procedures cause loss of useful information for the scientific studies. To prevent information-loss, — before releasing the information — users can be offered an identity separation feature, which, in a nutshell, means that people can present different information about themselves for their different groups of contacts. As these information can be contradictory sometimes, users may have more, independent identities in the same network. Although, such systems with partial identities are not so common yet, similar functions can be found in the social network Google+ and Facebook also starts to give the freedom to separate the information between the different groups of friends — only, these methods are not for profiles. The subjects of my work are these new kinds of networks: my goal is to show how we can derive a similarity measure, that allows us to examine the similarity between users, even in these networks which use identity separation techniques. To see how effective this derived measure is, the method will be compared with the traditional measures, and it will be examined in carefully generated data sets – as there are no such real networks available for scientific research yet.

Based on several articles in this field, I have found that for various purposes the most effective structural similarity measure is the *cosine similarity* ( $x, y$  are people in the network, while  $X, Y$  are their neighborhood-sets):

$$\text{Cos}(x, y) = \frac{|X \cap Y|}{\sqrt{|X| \cdot |Y|}}$$

In the second half of my thesis (after a brief introduction to de-anonymizing attacks, referral systems and structural similarity measures) I show some possible ways of deriving a new measure from the cosine similarity, then and analyze them. I compute the results for a smaller network (with 34 users) to visualize how these measures work, then make some simulations on my generated data sets, which have identity separation.

The results of the analysis can show, that identity separation alters the original similarity values of a network in a way, that can be useful for defending against the de-anonymizing attacks. Also, referral systems can use these measures to make more precise recommendations to users, based not only on their user profile, but also taking into account their different identities and how they separate the informations between the different groups of contacts.