Social Networking by Proxy: Analysis of Dogster, Catster and Hamsterster

Daniel Dünker
University of Koblenz–Landau
Universitätsstr. 1, 56070 Koblenz
dduenker@uni-koblenz.de

Jérôme Kunegis
University of Koblenz–Landau
Universitätsstr. 1, 56070 Koblenz
kunegis@uni-koblenz.de

ABSTRACT

Online pet social networks provide a unique opportunity to study an online social network in which a single user manages multiple user profiles, i.e. one for each pet they own. These types of *multi-profile networks* allow us to investigate two questions: (1) What is the relationship between the pet-level and human-level network, and (2) what is the relationship between friendship links and family ties? Concretely, we study the online pet social networks Catster, Dogster and Hamsterster, and show how the networks on the two levels interact, and perform experiments to find out whether knowledge about friendships on a profile-level alone can be used to predict which users are behind which profile.

1. INTRODUCTION

The specific topic that unifies an online community usually does not affect its basic mechanism: A user creates an account to connect with other users. Online pet social networks are however different in this regard, as they allow users to create any number of accounts, one for each pet they own.

In online pet social networks, a single user may (and is expected to) create one account for each owned pet. All social networking functionality such as entering personal information, creating friendship links to others, etc., are then performed on the pet level. With their structure that allows multiple profiles per account, online pet social networks thus make it possible to investigate the following questions:

- How does the fact that individual users own multiple profiles influence the structure of the social network?
- Is it possible to predict that two accounts are managed by the same person?

An overview of our datasets of Dogster.com, Catster.com and Hamsterster.com (defunct as of 2014) is given in Table 1. On all three sites, a single user can create accounts for any number of pets. Catster and Dogster are connected, and thus a single user account can be used for both sites. The group of pet profiles created by a single user makes up a *household* or *family*. Friendship links are allowed within a single household in Dogster and Catster, but are not allowed in Hamsterster. All friendship links are undirected.

2. HOMOPHILY IN PET NETWORKS

The term *homophily* refers to the tendency of people connected through social ties to be similar to each other. More precisely, homophily can be measured by a network's assortativity with respect to a given node property. A network then displays positive homophily (assortativity) when two randomly chosen connected persons are more similar than two randomly chosen persons without regard to connections [2]. Inversely, a network displays negative homophily (dissortativity) when the opposite is the case. By analysing the homophily in online pet social networks, we want to answer the following questions:

- Which is higher, the homophily between friends, or within families?
- Which profile properties correlate with two pets being friends, and with two pets being in the same household?

In order to answer these questions, we propose two complementary assortativity coefficients that apply to multi-profile social networks, whose ratio is measure of the relative strength of intra-household homophily as compared to across-friendship homophily.

Methodology. Many different node properties can be subject to homophily analysis, and the exact method used for measuring it depends on the data type considered. We define two measures of assortativity for multi-profile networks: one that measures homophily on the profile friendship level (r_{p}) and one that measures homophily on the account level (r_{a}). For the friendship level, we consider the friendship edges between pets in the networks. For the account level, we consider all pairs of pets that are in the same household. As in most social networks, we expect to observe a certain amount of homophily in the pet friendship network. We further hypothesize that the homophily between pets within a single household is larger than the homophily for pets connected by friendship links. Therefore, we compute measures of homophily for both levels, based on the available pet characteristics.

For categorical variables, we base the assortativity coefficients on [2, Eq. (2)]. The assortativity coefficients defined in this way equal one for perfect positive homophily, and lie between negative one and zero for negative homophily.[1] For

[1] The value cannot be exactly -1; see [2] for an explanation.

WWW 2015 Companion, May 18–22, 2015, Florence, Italy.
ACM 978-1-4503-3473-0/15/05.
http://dx.doi.org/10.1145/2740908.2745936 .

Table 1: Datasets analysed.

Dataset	#Pets	#Friendships	#Households	Pets per household
Catster	204,424	5,443,885	105,089	1.95
Dogster	451,710	8,543,549	260,390	1.73
Catster + Dogster	623,766	13,991,746	333,111	1.87
Hamsterster	2,950	12,531	1,575	1.87

numerical variables, we use the Pearson correlation coefficient between the numerical properties of connected pets, as defined in [2, Eq. (20)]. The values of this statistic range from -1 to $+1$ and are one for perfect positive homophily and -1 for perfect negative homophily. For the geolocation, we use the distance correlation [3] as a measure of homophily, based on the great circle distance between pairs of locations.

All three types of assortativity measures are zero when neither positive nor negative homophily is observed. To compare the both the assortativity coefficients on the friendship level and on the account level, we define the multi-profile assortativity ratio of a profile characteristic as $r_{\mathrm{rel}} = |r_{\mathrm{a}}/r_{\mathrm{p}}|$. By construction r_{rel} is larger than one if the assortativity is higher within profiles of one account than across friendships, and smaller than one if it is the assortativity across friendships that is higher.

Table 2: Results of family tie prediction.

	AUC			Regression weights		
Feature	**Cat**	**Dog**	**Ham.**	**Cat**	**Dog**	**Ham.**
Degree difference	82.3%	75.7%	72.3%	0.09	−0.27	0.22
Friend[a]	50.3%	50.6%	—	4.83	3.76	—
Common friends	79.0%	91.5%	71.7%	−0.46	0.71	4.98
Jaccard index	82.8%	92.2%	76.2%	5.78	9.73	1.25
Same race	66.4%	66.2%	76.4%	1.32	3.08	0.92
Same sex	51.9%	50.3%	54.2%	0.07	0.02	−0.09
Same coloration[b]	57.2%	—	59.4%	0.95	—	5.59
Same location	87.2%	90.3%	99.6%	11.02	8.92	21.21
Birth date difference	53.7%	50.1%	73.5%	−0.41	−0.30	0.42
Same join date	79.7%	74.6%	78.2%	6.08	5.44	6.21
Join date difference	90.8%	87.6%	91.9%	1.19	0.87	−0.24
Join age difference	52.7%	48.7%	66.2%	0.42	0.30	−0.88
Weight difference[c]	41.6%	—	—	−0.01	—	—
Same weight[c]	—	61.9%	—	—	0.52	—
Regression	99.3%	99.6%	99.9%			

[a] Hamsterster does not allow friendship links within one household.

[b] Dogster does not allow to specify a dog's coloration.

[c] Catster allows exact weights and Dogster has weight ranges.

3. PREDICTING FAMILY TIES

A family tie can be thought to exist between two pets that are in the same family, i.e., whose profiles were created by the same user account. We analyse in this section the task of predicting that two pets are in the same family, given only friendship links and pet-level profile metadata. This allows us to determine how well it can be predicted whether two profiles are from the same account, even when that information is not public. Since we have multiple types of profile data available, we can investigate which profile data allows to do this how well. Also, the experiment serves to find out which properties of pets are consistent within a household, and which are independent of a household.

Prediction Methods. Given a multi-profile social network $G = (V, W, E, m)$, we want to predict whether two profiles are managed by the same account, i.e., information contained in W and m, using only the profile-level network $G_{\mathrm{p}} = (V, E)$, including the metadata associated with it. In the case of pet social networks, we use the available pet profile information along with the pet-level friendship links for learning. We investigate the following indicators (i.e., features), each of which applies to a pair of profiles $\{u, v\}$. We do not use geographical distance between the two profiles, because we know that if the distance is larger than zero, then the profiles must be in distinct households. Thus, we only use the "same location" feature. Note also that the geolocation is given only up to the city level, i.e. all pets in New York City will be counted as having the same location, leading to a large number of pets from different households but with the exact same location. We also perform a logistic regression prediction, combining all features given above.

Experiments. In order to measure the accuracy of each prediction method, we use a test set defined in the same manner as the training set, i.e., we randomly sample e pet pairs known to be in the same family, and e pet pairs known not to be in the same family. This test set is disjoint from the training set used for learning the regression parameters. The accuracy of the prediction methods is measured using the area under the curve (AUC) [1]. The AUC is 1/2 for a random prediction, and one for a perfectly accurate prediction. Table 2 gives the AUC values for each method separately and for the regression predictions, as well as the learned regression weights for each of the three sites.

Discussion. We observe that in all three sites, pets in the same household can be detected with an AUC of over 99% using the regression predictor. This means that given two pairs of pets, one of which from the same household and one of which from two different households, our algorithm will detect which is which in over 99% of cases. This high value can be explained by the fact that certain individual indicators are already highly indicative of family ties.

4. REFERENCES

[1] A. P. Bradley. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159, 1997.

[2] M. E. J. Newman. Mixing patterns in networks. *Phys. Rev. E*, 67:026126, 2003.

[3] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *The Ann. of Statistics*, 35(6):2769–2794, 2007.