Combining Propensity and Influence Models for Product Adoption Prediction

Ilya Verenich, Riivo Kikas, Marlon Dumas, Dmitri Melnikov
University of Tartu, Estonia
{ilyav, riivokik, marlon.dumas, zx}@ut.ee

***Abstract*—This paper studies the problem of selecting users in an online social network for targeted advertising so as to maximize the adoption of a given product. In previous work, two families of models have been considered to address this problem: direct targeting and network-based targeting. The former approach targets users with the highest propensity to adopt the product, while the latter approach targets users with the highest influence potential – that is users whose adoption is most likely to be followed by subsequent adoptions by peers. This paper proposes a hybrid approach that combines a notion of propensity and a notion of influence into a single utility function. We show that targeting a fixed number of high-utility users results in more adoptions than targeting either highly influential users or users with high propensity.**

I. Introduction

A central problem in modern marketing is that of constructing decision models to select potential customers to target in a marketing campaign in such a way as to maximize the ensuing number of adoptions. A *direct marketing* method consists in selecting those individuals from a potential customer population who have higher propensity to respond positively to the campaign, e.g. higher propensity to adopt a given product [1]. In the context of online social networks, this method requires building a decision model that predicts the response of each individual given the available data such as geography, demographics and past behavior of the user in question and their peers. The direct marketing approach does not take into account that an individual's adoption may have an effect on adoption by others [2]. However, some markets – most notably those associated with information goods – exhibit strong network effects [3], meaning that individuals are often strongly influenced by their peers. An alternative *network-based marketing* approach takes advantage of this property by targeting primarily individuals with the strongest influence potential. In this approach, the decision model selects individuals who are likely to influence one or more of their peers into adopting the product (i.e. "word-of-mouth" effect).

In this paper, we present a method for constructing predictive models that combine the notions of propensity and influence in order to identify likely product adopters in a communication network. We first develop, as a baseline, an adoption propensity model based on previous work [4]. Second, as our dataset is missing explicit product diffusion paths, we propose a method to infer influence from the underlying network of interpersonal communications and the temporal sequence of product adoptions. Third, we develop a model to estimate user influence based on the inferred influence links. The proposed influence prediction model specifically identifies individuals whose adoption is likely to trigger at least one subsequent adoption among their friends. Finally, we define a combined model that brings together the notions of propensity and influence into a single utility function. The proposed models are comparatively evaluated using a dataset of a global communication network, namely Skype. The results show that the combined model provides significant improvement relative to the separate models based on propensity or influence alone.

The rest of the paper is structured as follows. Section II introduces the dataset used in the study and the features extracted thereof for product adoption prediction. Next, Section III presents the three proposed models, while Section IV discusses their evaluation with respect to accuracy and marketing effectiveness. Finally, Section V analyzes related work, while Section VI provides concluding remarks.

II. Datasets and features

This section provides an overview of the dataset used in this study, and the features used for constructing the adoption prediction models.

A. Dataset description

The study has been conducted on a dataset of the Skype social network. The centerpiece of the dataset is the evolving *contact network*, where the nodes represent users, and there exists an edge between a pair of users if they are in each other's contact lists. A user's contact list is composed of a user's *friends*. If a user u wants to add another user v in their contact list, u sends v a contact request, and the edge is established at the moment v approves the request (or not established if the contact request is not approved). Each edge is labeled with a timestamp indicating a moment the contact request was approved. The dataset includes circa 450 million users and 3 billion edges.

Every user has a set of demographic and geographic attributes that can be optionally filled in their profile. Three such attributes are present in the dataset – gender, birth year and country. However, users may leave any of these three fields blank. In addition, for each user, the dataset indicates the Skype client platform from which the user last connected (e.g. Windows client, Mac OS client, iPhone client), and a number of attributes that are automatically filled upon the user's registration: profile creation date, country code and location code where profile was created. Country and location information is originally extracted via IP address geocoding when the user registers their account.

In Skype, users can chat with each other or make audio or video calls free of charge. In this respect, the dataset

TABLE I: Model features.

Set	Feature	Description
Topological features	*fr*	Number of friends
	sf	Number of product friends
	sfr	Product friends ratio
	foc	Number of friends in other countries relative to *fr*
	fol	Number of friends in other locations relative to *fr*
	sfc	Number of product friends in other countries divided by *fr*
	sfl	Number of product friends in other locations divided by *fr*
	ccf	Clustering coefficient of a user's egocentric network
Temporal features	*fr2*	Number of friends added during the last 2 months
	sf2	Number of product friends during the last 2 months
	dccf	Absolute change in *ccf* during the last 2 months
	longev	Average length of acquaintance with a user's friends
Profile features	*age*	Age group
	gend	Gender
	accAge	Account age
	plat	User's platform or operating system
	country	User's country code
Usage intensity features	*avgCon*	Average connected days per month
	lastCon	Number of connected days last month
	AvgCht *AvgAud* *AvgVid*	Weighted average percentage of instant messaging, audio call and video call days per month
	lastCht *lastAud* *lastVid*	Number of instant messaging, audio call and video call days last month

includes for every user and for every month since the user's account creation, the number of days in the month when the user chatted, audio-called or video-called. Usage data is not available at a lower granularity.

In addition to the above free services, users can purchase "credits" for calling phones or to send SMS messages (among other purposes). The dataset includes for each user, the date when the user first adopted each of two paid products: *Buy Credit* (first credit purchase, for any purpose) and *SMS* (first SMS sent). Herein, these are called *product adoption events*.

The dataset does not include identity information. All usernames are anonymized and there is no means to infer a user's identity solely from their profile – location data is only available at a granularity where there are at least thousands of users per location. The dataset does not include any information about interpersonal interactions, besides the fact that a user is in another user's contact list.

B. Feature description

For the purposes of constructing (predictive) classification models, each individual in the network is abstracted as a set of features. Below, we describe and motivate the features we extract from each user based on their own attributes and history, and those of their immediate social network (also called the egocentric network). All features are listed in Table I, grouped into four categories.

1) Topological features: Topological features capture the structure of the network of interpersonal relations. These features are well-known and have been extensively studied in different contexts [4]–[6]. Number of friends, abbreviated as *fr* in Table I, is the simplest network feature that is computed by counting the number of contacts in the user's contact list. In our study we discard users whose contact list is empty, as it is not possible to compute network features for them.

The number of network neighbors who already use a product has been proven important for estimating the probability of adoption [4], [5], [7], [8], thus suggesting the presence of peer pressure effect. For convenience, we will refer to a user's friends who have adopted the product as product-using friends, or simply *product friends*. Consequently, we define a feature *sf* that is equal to the number of product friends in the neighborhood of a particular user. We also define product friends ratio $sfr = sf/fr$ as the ratio of the number of product friends to the total number of a user's friends.

With a large fraction of friends in other countries or locations, a user may find that paid products provide an easier and more convenient way to communicate with friends abroad or far away. Perhaps after migration to another area a user feels the need to stay in touch with their relatives. Therefore, we decided to include features *foc* and *fol*, calculated as the fraction of user's friends whose country or location is different from their home country or location respectively. Features *sfc* and *sfl* are computed analogously, except we count only *product* friends in other countries and locations.

We also take into account local clustering coefficient (ccf) as it is known that higher clustering coefficient favors propagation of products in the networks, since nodes tend to be more tightly connected [9].

It should be noted that since the network is evolving, topological features change over time and thus need to be computed at the time that predictions are made (cf. Section III-A).

2) Temporal features: Temporal features reflect the change in the neighborhood of an individual over time. Such features have been extensively studied in the domain of dynamic networks [10], [11]. With the inclusion of these features we try to capture possible dynamic process happening in the user's network just before the product adoption.

One of the simplest temporal features is user's dynamic degree, counted as the rate, at which new friends are gained [11]. The importance of dynamic degree for information diffusion in the network has been acknowledged by Luu et al. [12]. In our study we approximate dynamic degree by calculating number of friends a user has added during the last two months (*fr2* in Table I). We also count number of friends who adopted the product during the last two months and denote it as *sf2*. Analogously we approximate dynamic clustering coefficient ($dccf$) [11] as the absolute change in the clustering coefficient of user's egocentric network over the last two months.

We also include the length of acquaintance as one of the indicators characterizing strength of interpersonal ties [13]. It is natural to assume that two individuals tend to share higher "level of trust" if they know each other for a longer period of time. Since only a small fraction of trusted friends has the real influence on a user [14], determining such trusted friends by their length of acquaintance can benefit the model. In this study we calculate the average length of acquaintance of a user with their neighbors, expressed in months (*longev*).

3) Profile features: Profile features are taken from users' account description. These features carry demographic and geographic information and usually do not change over time. Aral and Walker [15] provided insights into how demographic parameters, such as age, gender, relationship status, affect personal influence and susceptibility towards product adoption.

We include user's age, gender, country of registration and their Skype client platform as basic profile features. Additionally we calculate account age as the time elapsed since a user created a network account. Introduction of this feature will allow us to distinguish users who created an account specifically for using paid products. Previously, Thompson and Sinha [16] showed that community membership duration affects the likelihood of adopting a new product.

4) Usage intensity features: Previous research has indicated that in online communities which combine open and proprietary products or services, as consumers climb up the "ladder of engagement", they develop a deeper sense of commitment to the website [17] and perceived ownership [18]. Oestreicher-Singer and Zalmanson [19] in their study on Last.fm network also discovered that the more active a user is, the more likely they are to adopt a paid product. With this intuition, we extract a set of features describing intensity of usage of the other products – instant messaging (chat), audio- and video. Intuitively, we expect users that are active with some products to be also active with the other ("target") products.

Features $lastCht, lastAud$ and $lastVid$ show how many days chat, audio- and video calls respectively were used during the latest month. These features are based on the assumption that adopters increase their activity in the month before adopting. In addition, feature $lastCon$ shows the number of days a user connected to the network during the last month, and is used to filter out inactive (dormant) users. It is unlikely that such users will suddenly adopt the product. Similarly, we define $AvgCht$, $AvgAud$ and $AvgVid$ as the average number of days a particular free product has been used in the past, starting from either the time a user has created account or the time of the data recording.

III. MODELS

In this section we discuss the construction of the three models for the product adoption.

A. Adoption propensity model

As mentioned in the introduction, a central task in direct marketing is to identify users who are the most likely to adopt a given product. This task can be recast as a ranking problem: given a set of users V and their features, rank them according to their probability P_u $(u \in V)$ to become an adopter during a certain time period. For convenience, we will refer to the estimated probabilities P_u as adoption propensity scores, or simply *propensity scores*.

Given that this is a predictive task, we apply a temporal split to the dataset. Specifically, we fix a time point T_1 as the moment when the prediction is made. We use data from a past interval (T_0 to T_1) for training, and data from a future interval (T_1 to T_2) for testing. Users who adopted between T_0 and T_1 are positive examples and their features are calculated at the time of their adoption. Users who do not adopt during this period are negative examples and their features are computed at T_0. Users who had adopted the product prior to time T_0 are excluded.

For every user in the test set, features are computed at time point T_1. Users who adopt the product between T_1 and T_2 are the positive examples and all others are negative examples.

B. Inferring influence links

We have noted that network-based marketing targets individuals who are likely to trigger further product adoptions. Thus, a decision model to support this type of campaign should be able to pinpoint individuals who will "influence" others into adopting the product in question. Some social networks capture explicit links of influence (or diffusion) between individuals, for example in the form of retweets and mentions [6], reshares [20], recommendations [21], etc. In this paper however we deal with a social network that does not capture explicit influence links. Thus, we need to infer these links from the network of interpersonal connections and the temporal sequence of adoptions [22].

To test the presence of interpersonal influence in our network, we calculate the distribution of product adoption inter-event times, i.e. the time between any pair friends adopting the product, in the case where the link between them was created before the first of them adopted. Additionally, we calculate the inter-event time between all possible pairs of adopters, regardless of whether they are friends or not. The probability density function (PDF) of the adoption inter-event time among pairs of friends – shown by the red line on Fig. 1 – indicates a decaying behavior. In other words, when a friend of a user adopts the product, their likelihood of adoption is higher than random, and this difference decays over time. A similar distribution was found by Goyal et al. in their study of the Flickr network [23].

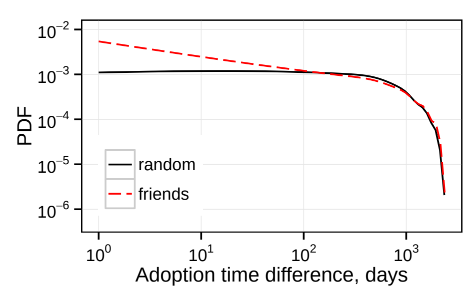


Fig. 1: Adoption time difference between pairs of adopting friends and pairs of random adopters. This plot is for the *Buy credit* product. The plot for the *SMS* product is similar.

Fig. 1 also shows that beyond an interval of approximately 90 days, the probability of a pair of friends (u, v) adopting after each other is similar to the probability of two random (possibly unrelated) pairs of users (u, v) adopting after each other. In other words, beyond 90 days there is no temporal correlation (beyond chance) between subsequent adoptions by a pair of friends. Accordingly, we assume that an influence link exists from user u to user v when:

- v adopted the product after u and within $\Delta t = 90$ days
- $(u, v) \in E$ was created before v adopted

A similar approach is used in Dave et al. [24] to separate between social influence and homophily in the context of pairs of friends performing the same action.

C. User influence model

Having defined influence and determined the temporal threshold Δt, we note that a possible measure of user influence could be the number of subsequent adoptions N within Δt days since user's own adoption. However, this variable is very skewed – after around 82% of adopters no subsequent adoption happens in their neighborhood for Δt days, around 14% of adopters are followed by one further adoption, and the long tail (2 to over 30 adoptions) accounts for less than 4% of adopters (Fig. 2). Normalizing the number of adoptions by the number of user friends *fr* does not solve the problem, as for 15% of adopters $0 < N/fr < 0.2$ and for 3% adopters $0.2 \leq N/fr \leq 1$.

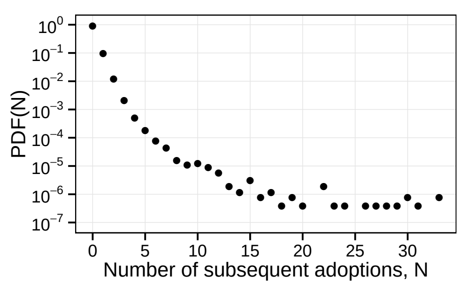


Fig. 2: Distribution of the number of subsequent adoptions within $\Delta t = 90$ days.

Thus, instead of predicting the number of subsequent adoptions, we will simply predict whether a user's adoption will be followed by any of their neighbors' adoption within Δt days. We can formulate this task as a ranking problem: given a set of users V, who will presumably adopt the product, and their features, order them according to their probability I_u ($u \in V$) to trigger subsequent adoptions in their neighborhood within Δt days.

For convenience, we will refer to I_u as *influence scores*, and users for whom $I_u > 0.5$, i.e. users who are predicted to trigger at least one subsequent adoption, *influential users*.

To solve this problem we use the same set of features as in the previous model. However, there are some differences in creation of training and test sets. For the training set we take all users who adopted the product from T_0 to T_1. Those adopters with at least one subsequent adoption within Δt days after their own adoption are positive examples. The negative examples are all other users with no subsequent adoption in their neighborhood. For all examples features are computed by the time of their adoption.

For the test set, we take the users who adopted the product from T_1 to T_2. For every user in the test set, features are also computed at the time of their adoption. As a result of running the classification algorithm we would like to put every influential adopter to the top of our ranked list of users.

D. Utility-based model

In the viral marketing campaign an advertiser aims to find the optimal group of the most profitable customers to target in order to trigger the widespread adoption of a new product or innovation. To account for these two criteria, we apply the framework developed by Domingos and Richardson [2], [3] who modeled a consumer network as a Markov random field for maximizing profit. They distinguished between a customer's *intrinsic value*, which derives from the purchases they will make, and *network value*, which derives from their influence on other customers. The authors tested their model on a database of movie reviews and found that their proposed methodology outperforms non-network methods for estimating customer value.

Let c be the cost of marketing to a user u (assumed constant), P_u be the u's propensity score, i.e. probability of purchasing the product, S be the unit price of the product, and M_u be the amount of product, consumed by the user u. Since in our study we focus on one product, which can be adopted or not, a user's intrinsic value J_u can be determined as

$$J_u = P_u S M_u - c. \tag{1}$$

However, for the paid products we only know the dates of the first and last product usage (see Section II-A), from which we cannot infer the usage intensity. Therefore, we assume everyone who has adopted the product, would use it in an equal amount (for convenience, we set it to one unit):

$$J_u = P_u S - c. \tag{2}$$

The network value N_u of a user u is high when they are expected to have a very positive impact on others to purchase the product (e.g., through word of mouth). Consequently, N_u is proportional to the number of subsequent adoptions A_u user u triggers after their own adoption:

$$N_u = P_u A_u S. \tag{3}$$

It should be noted that such subsequent adoptions can be triggered only if user u adopts the product. However, a marketer, when targeting users, does not know who will indeed adopt and who will not.

Combing intrinsic and network values, we define a total value T_u of a user u, or user's *utility* as:

$$T_u = J_u + N_u = P_u S(1 + A_u) - c. \tag{4}$$

With our dataset, we cannot accurately predict the total number of subsequent adoptions A_u, since its distribution is very skewed (Fig. 2). Instead, we trained a classifier to estimate user's influence score I_u, i.e. probability that their adoption will be followed by any of their friends. which correlates with A_u with the Pearson correlation coefficient 0.415 ($P < 10^{-6}$, 95% CI 0.414 to 0.416). Thus, we can rewrite user's utility as:

$$T_u = P_u S(1 + I_u) - c. \quad (5)$$

We hypothesize that targeting users with higher utility will result in more adoptions than targeting either users with higher influence score or higher propensity score. The intuition behind this comes from the fact that we observe no significant correlation between users propensity P_u and influence scores I_u. The only noticeable exception is users (less than 1% of all adopters) with high influence, who tend to have high propensity to adopt. However, the opposite is not necessarily true – users who are almost surely to adopt the product may still have near-zero influence.

To validate our hypothesis, we sample a set of 10 million users, to which we will refer as V. For every user in V, we calculate their propensity and influence scores with the two previous models. Then we calculate the utility scores (Equation 5) and order users according to them. We count how many product adoptions occurred among top $X\%$ of the ordered users during the next six months. In case a user adopted the product, we count how many subsequent adoptions happened in their neighborhood for the following 90 days.

To calculate utility scores we use Equation 5. Since parameter S and c are equal for all users, they will not affect user ranking. For convenience, we set S to 1, and c to 0. The resulting utility distribution shows most users have low utility score. Specifically, for less than 7% of users $1 < T_u \leq 2$. A similar distribution was observed by Domingos and Richardson [2], [3].

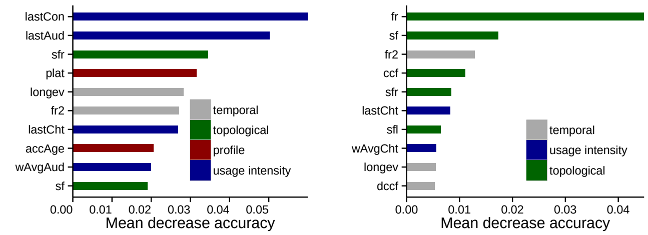
IV. Evaluation

To train the classifiers for the adoption propensity and user influence models, we use random forest, as implemented in the GNU R package `randomForest`. For each model we train 500 trees, while keeping the default value of the number of variables randomly sampled for each tree $m = \lfloor \sqrt{M} \rfloor = 5$. In this section we evaluate their performance on the test set. Obtained propensity and influence scores serve as input for the combined utility-based model.

The training interval (T_0 to T_1) is fixed in the evaluation and corresponds to a period of one year in the past. Parameter T_2 was varied so that the test period spans 3, 6, 9 and 12 months. Below we only report results for $T_2 - T_1 = 6$ months. The accuracy observed for 3-months test periods was slightly higher but within two percentage points of the accuracy for 6 months test periods. Similarly, the accuracy observed for 9 and 12-months test periods was slightly lower but also within two percentage points of the accuracy for 6-months test periods. In all cases, the relative accuracy (gain) of the models remains the same for different prediction time windows.

A. *Propensity and Influence Models*

Fig. 4a plots the cumulative gains chart of the propensity-based model applied to both *Buy credit* and *SMS*. The diagonal



(a) propensity (b) influence

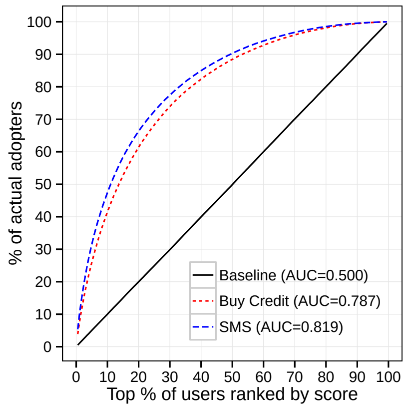
Fig. 3: Top 10 important features according for the propensity and influence models, measured by mean decrease accuracy.

in the chart corresponds to the performance of a random model that assigns all the users random uniform probabilities from 0 to 1 to adopt the product. A point in the cumulative gains chart plots the percentage of actual adopters included in the top-X% of the population ranked by propensity score (adoption probability). For example, we see that for the *Buy Credit* product, the top 10% of the population ranked by propensity score contains around 41% of all the adopters. Meanwhile, for *SMS*, the top 10% of the population contains 47% of all the adopters of this product. The figure also indicates the Area Under the Cumulative Gains chart (herein AUC), which provides an aggregate measure of accuracy. We observe that the prediction accuracy for *SMS* is slightly higher compared to the *Buy Credit* product, but not significantly – particularly not beyond the top-10 percentile of the population.

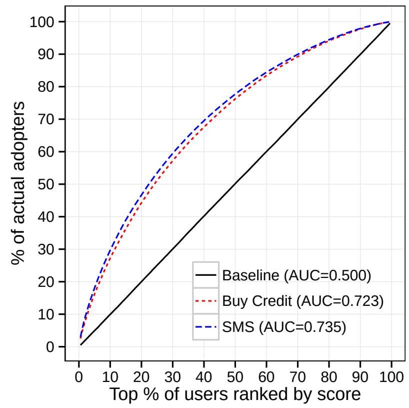
The cumulative gain chart of the influence model is given in Fig. 4b. The y-axis in this chart gives the percentage of all "influential adopters" included in the top-X% of the population ranked by influence score – where an influential adopter is an adopter whose adoption was followed by at least one other adoption within Δt. We observe that for both products, the influence model has lower predictive power than the adoption propensity model. For example, we see that random forest can order the test set in such a way that the top 10% would contain around 28% of all the adopters of the *Buy Credit* and 30% of all the adopters of the *SMS*. This observation suggests that the overall effect of influence is less strong than that of propensity.

In order to shed light into the features responsible for the observed predictive accuracy, Fig. 3 shows the relative importance score – measured via *Mean Decrease Accuracy* – of the top ten features for each of the two models and for the *Buy Credit* product. We note that very similar results are obtained for the *SMS* product. In the case of the propensity model, Fig. 3a shows that the most predictive features of user adoption are *lastCon* and *lastAud*. Thus, the activity of the user in the month prior to adoption is a good indicator of a potential future adopter. Product friends ratio, *sfr*, is also among the most important features, which indicates the presence of peer-pressure effects in the network.

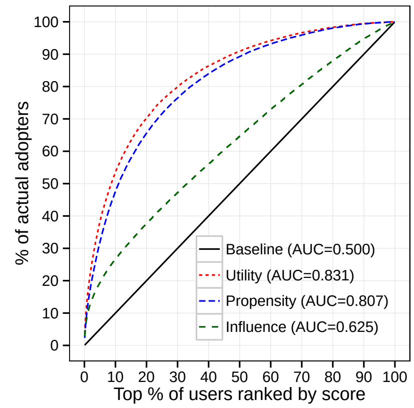
On the other hand, the most predictive feature for user influence is the number of friends *fr* (Fig. 3b), thus confirming previous studies that observed the importance of centrality-based measures for content diffusion [25]. One can indeed argue that having more friends increases the probability that



(a) Propensity



(b) Influence



(c) Utility

Fig. 4: Model performance

at least one of them will adopt within fixed time Δt. In fact, retraining the model with only one feature *fr* gives about a half of the observed prediction accuracy.

B. Utility-based model

The utility-based models aim at identifying adoptions both by the selected users and their friends. Accordingly, to evaluate this type of model via cumulative gains charts, we consider in the y-axis both *primary adoptions* A_I and *subsequent adoptions* A_{II} within the top-X% of users in the population. Primary adoptions refer to users in the top-X% of the population who actually adopted, while subsequent adoptions refer to users who are friends of a user u in the top-X% of the population and who adopted after u within the window $\Delta t = 90$ days. Importantly, we count only unique adoptions. For example, suppose $u_1 \in V$ adopted the product at time t_1 and $u_2 \in V$ adopted at time t_2, such that $T_0 \leq t_1, t_2 \leq T_1$, and u_1 is ranked higher than u_2. If $u_3 \notin V$ had been a friend of u_1 by the time t_1 and a friend of u_2 by t_2, and u_3 adopted at t_3, such that $0 < t_2 - t_1 \leq \Delta t$ and $0 < t_3 - t_1 \leq \Delta t$, then u_3's adoption is only counted once.

Fig. 4c shows the resulting cumulative gains chart with three curves (besides the diagonal) obtained based on the rankings by propensity score P_u, influence score I_u and utility score T_u respectively, using the data for the *Buy Credit* product. The chart also provides the corresponding AUC scores for each curve. The chart shows the utility-based model outperforms the propensity one by a small but visible margin. For example, targeting 10% of users from set V, ordered by utility score, produces 53.5% of all adoptions that would happen in the set, including subsequent adoptions in their neighborhood. The same fraction of users ranked by propensity score produces 47.5% of adoptions, and by influence score this number drops to 26.7%. For the *SMS* product the AUC values are within three percentage points of the specified values in Fig. 4c. It should be noted, however, that most gain comes from the propensity component, since the number of primary adoptions is much higher than the number of secondary adoptions.

TABLE II: Number of acquired paid users as a function of the number of targeted users.

# of users targeted	# of acquired users, when ordering by:		
	Propensity	Influence	Utility
100	8	**67**	53
500	78	**217**	187
1000	190	332	**335**
10000	1184	1464	**1839**
100000	6028	4871	**7847**
1000000	24242	13590	**27274**

So far, we have evaluated the models in terms of their accuracy measured on the basis of their cumulative gains chart. In the context of targeted advertising, another common approach to evaluating a decision model is based on the amount of predicted adoptions when targeting a fixed-size population. Along this line, Table II shows the absolute numbers of *acquired users* as a function of the number of *targeted users* T, where the set of targeted users is determined by taking the T top-ranked users in order of adoption likelihood according to a given model. For each T, the method with the highest number of acquired users is shown in bold.

The following observations can be made:

- Targeting a fixed amount of users generally results in higher amount of adopters for the same period of time, if we order them by *utility* score. The exact improvement depends on the number of targeted users T and the baseline (propensity- or influence-based ordering).
- If T is less than about 1000, or 0.01% of the network population, ranking users by *influence* score is the optimal decision.
- If T is less than about 30000, or 0.3% of the network population, ranking by *influence* score is better than ranking by *propensity* score.

The last two observations can be explained by two factors. First, we observe that users with high (say > 0.8) influence score tend to have high propensity score as well, but the reverse is not necessarily true: even if the user has high propensity score, they may still have near-zero influence score. Therefore, when targeting highly influential users, we are also targeting users who are most likely to adopt. In this way, we select individuals with high intrinsic value and high network value. Second, since influence score is moderately correlated with the number of subsequent adoptions, by taking users from the top of the list ordered by influence score, we capture those who are followed by many adoptions, and therefore contribute to the total number of adoptions at a faster rate. However, as only less than 4% of adoptions are followed by two or more adopting friends, ranking by influence score loses its advantage as we choose more users to target, first to ranking by utility ($T > 0.01\%$ of the network), then to ranking by propensity ($T > 0.3\%$).

Finally, to check the robustness of the above observations, we repeated the whole procedure twice, randomly sampling sets of users of the same size and under the same conditions. The results are similar to the previously discussed. Specifically, the AUC across different experiments stays within 1.5 percentage points of the values provided in the Fig. 4c, and in all cases the highest AUC value is achieved with the utility-based model.

V. Related work

An extensive amount of research has been done in both online and offline social networks to understand and quantify social behavior, information diffusion and mechanisms of product adoption.

Perhaps the most relevant work to ours is by Bhatt et al. [4], who studied the spread of the *PC to Phone* product in a network, providing communication services. They found that the spread of product adoption is not so much due to the presence of individual influencers, but is rather a result of influence yielded by peer-pressure where users with more adopter friends were more likely to adopt themselves. They also showed that the model combining both *user* and *social* features to estimate product adoption propensity is more accurate that models that use either user or social features in isolation. This work however focuses exclusively on propensity and does not consider influence.

Other studies have provided evidence of "peer pressure" effects in social networks. For instance, Hill et al. [5] analyzed marketing campaign data of a large telecommunications company and found that consumers linked to prior customers are themselves more likely to adopt the product. Sundsøy et al. [7] found that probability of adoption *iPhone* is proportional to the number of adopting friends. Liu and Tang [8] also discovered that a user is more likely to adopt if the product has been widely adopted by their friends. These observations underpin the choice of features in the proposed propensity model.

Another body of related work focuses on identifying influential individuals and studying their role in the process of diffusion of innovation. Considering high-degree nodes as influential, known as degree centrality, has long been a standard approach [26]. This has been proven true in our case, as well (Fig. 3b). In contrast, Onnela and Reed-Tsochas [27] found that high-degree users are not necessarily the source of influence and that only a small fraction of their friends adopt after them. Bughin et al. [14] discovered that it is the small, close-knit network of trusted friends that has the real influence on a particular user. Iyengar et al. [28] discovered that the amount of interpersonal influence is moderated by both the recipients' perception of their opinion leadership and the sources' volume of product usage. Cha et al. [6] in a study of a Twitter dataset, discovered that a high follower count does not always lead to many retweets and mentions.

Hinz et al. [29] studies the product adoption by means of social influence in friendship-based networks, such as Skype and advice-based networks which include topical subnetworks, such as Google+. They conclude that only advice-based networks clearly identify influential individuals.

Watts and Dodds [30] contemplate that large cascades of influence are driven not by influentials but by a critical mass of easily influenced individuals. Davin et al. [31] also challenge the influence hypothesis, arguing that latent homophily could inflate the proportion of adoptions attributed to social influence by 40% and in some samples by over 100%. Shalizi and Thomas [32] show that homophily and social influence are generally confounded with each other; thus, distinguishing between them requires strong parametric assumptions. In our case to assert the notion of influence and separate it from the homophily, we used the temporal threshold Δt.

VI. Conclusion

This study has put into evidence the inherent complementarity of propensity-based and influence-based models for predicting product spread in a large-scale communication network.

First, we have shown that a propensity model combining past user behavior, demographic and network features can achieve relatively high levels of accuracy (AUC in the order of 80%). Second, we have put into evidence the effect of influence in the dynamics of product adoption and derived influence links via temporal correlation, which then allow us to build an influence-based model for product adoption prediction. While this latter model is not as accurate (AUC in the order of 73%), we have then shown that the influence-based model can be combined with the propensity-based one into a single model that outperforms the two models separately. Moreover, we have shown that when cast in the context of targeted advertising campaigns with a fixed number of targets, the combined model generally leads to higher numbers of identified adoptions (i.e. customer acquisitions).

There are several potential extensions that could be incorporated into our model in order to increase its predictive power. First, we modeled user influence as a rectangular function that is non-zero during a given time range starting from the moment a user adopts the product. It may be possible however and potentially advantageous to model influence as a decay function, which would be in line with the observed distribution of inter-adoption times between friends. Second, when predicting subsequent adoptions attributable to influence, we did not take into account the users' own propensity to adopt the product independently of the influence effect. Taking into account this influence-independent propensity might lead to a more accurate influence-based model. Third, we could apply ensemble methods (particularly stacking) in order to find the optimal weights to assign to the influence and propensity scores when constructing the utility-based model.

Acknowledgments. This research is supported by Microsoft/Skype and ERDF via the Software Technology and Applications Competence Centre (STACC). The authors acknowledge the valuable input and comments of Ando Saabas and Adriana Dumitras.

REFERENCES

[1] G. Lantos, *Consumer Behavior in Action: Real-Life Applications for Marketing Managers*. M. E. Sharpe Incorporated, 2010. [Online]. Available: http://books.google.ee/books?id=JemkYebV5NYC

[2] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2001, pp. 57–66.

[3] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002, pp. 61–70.

[4] R. Bhatt, V. Chaoji, and R. Parekh, "Predicting product adoption in large-scale social networks," *ACM Conference on Knowledge Discovery and Data Mining*, pp. 1039–1048, 2010.

[5] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, pp. 256–276, 2006.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," *Proceedings of International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2010.

[7] P. R. Sundsøy, J. Bjelland, G. Canright, K. Engø-Monsen, and R. Ling, "Product adoption networks and their growth in a large mobile phone network," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*. IEEE, 2010, pp. 208–216.

[8] K. Liu and L. Tang, "Large-scale behavioral targeting with a social twist," in *Proceedings of the 20th ACM international conference on Information and knowledge management*. ACM, 2011, pp. 1815–1824.

[9] M. Cha, A. Mislove, and K. P. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proceedings of the 18th international conference on World wide web*. ACM, 2009, pp. 721–730.

[10] Y. Yu, T. Y. Berger-Wolf, J. Saia *et al.*, "Finding spread blockers in dynamic networks," in *Advances in Social Network Mining and Analysis*. Springer, 2010, pp. 55–76.

[11] H. Habiba, "Critical individuals in dynamic population networks," Ph.D. dissertation, Northwestern University, 2013.

[12] D. M. Luu, E. P. Lim, T. A. Hoang, and F. C. Chua, "Modeling diffusion in social networks using network properties." in *International AAAI Conference on Weblogs and Social Media*, 2012.

[13] P. V. Marsden and K. E. Campbell, "Measuring tie strength," *Social forces*, vol. 63, no. 2, pp. 482–501, 1984.

[14] J. Bughin, J. Doogan, and O. J. Vetvik, "A new way to measure word-of-mouth marketing," *McKinsey Quarterly*, vol. 2, pp. 113–116, 2010.

[15] S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," *Science*, vol. 337, no. 6092, pp. 337–341, 2012.

[16] S. A. Thompson and R. K. Sinha, "Brand communities and new product adoption: The influence and limits of oppositional loyalty," *Journal of marketing*, vol. 72, no. 6, pp. 65–80, 2008.

[17] P. J. Bateman, P. H. Gray, and B. S. Butler, "Research note – the impact of community commitment on participation in online communities," *Information Systems Research*, vol. 22, no. 4, pp. 841–854, 2011.

[18] J. Preece and B. Shneiderman, "The reader-to-leader framework: Motivating technology-mediated social participation," *AIS Transactions on Human-Computer Interaction*, vol. 1, no. 1, pp. 13–32, 2009.

[19] G. Oestreicher-Singer and G. Zalmanson, "Paying for content or paying for community? the effect of social computing platforms on willingness to pay in content websites," Working paper, Tel-Aviv University, Tech. Rep., 2011.

[20] P. A. Dow, L. A. Adamic, and A. Friggeri, "The anatomy of large facebook cascades," in *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media, ICWSM*, 2013.

[21] J. Leskovec, A. Singh, and J. Kleinberg, "Patterns of influence in a recommendation network," in *Advances in Knowledge Discovery and Data Mining*. Springer, 2006, pp. 380–389.

[22] G. Sharad, D. J. Watts, and D. G. Goldstein, "The structure of online diffusion networks," *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 623–638, 2012.

[23] A. Goyal, F. Bonchi, and L. V. Lakshmanan, "Learning influence probabilities in social networks," in *Proceedings of the 3rd ACM international conference on Web search and data mining*. ACM, 2010, pp. 241–250.

[24] K. Dave, R. Bhatt, and V. Varma, "Modelling action cascades in social networks," in *International AAAI Conference on Weblogs and Social Media*, 2011. [Online]. Available: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2741

[25] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2009, pp. 199–208.

[26] S. Wasserman, *Social network analysis: Methods and applications*. Cambridge university press, 1994, vol. 8.

[27] J.-P. Onnela and F. Reed-Tsochas, "Spontaneous emergence of social influence in online systems," *Proceedings of the National Academy of Sciences*, vol. 107, no. 43, pp. 18 375–18 380, 2010.

[28] R. Iyengar, C. Van den Bulte, and T. W. Valente, "Opinion leadership and social contagion in new product diffusion," *Marketing Science*, vol. 30, no. 2, pp. 195–212, 2011.

[29] O. Hinz, C. Schulze, and C. Takac, "New product adoption in social networks: Why direction matters," *Journal of Business Research*, vol. 67, no. 1, pp. 2836–2844, 2014.

[30] D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," *Journal of consumer research*, vol. 34, no. 4, pp. 441–458, 2007.

[31] J. P. Davin, S. Gupta, and M. J. Piskorski, "Separating homophily and peer influence with latent space," *Available at SSRN 2373273*, 2013.

[32] C. R. Shalizi and A. C. Thomas, "Homophily and contagion are generically confounded in observational social network studies," *Sociological methods & research*, vol. 40, no. 2, pp. 211–239, 2011.