Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)

The Role of Optimal Distinctiveness and Homophily in Online Dating

Danaja Maldeniya
University of Michigan
dmal@umich.edu

Arun Varghese
University of Michigan
arunv@umich.edu

Toby Stuart
University of California, Berkeley
tstuart@berkeley.edu

Daniel M. Romero
University of Michigan
drom@umich.edu

Abstract

Users of online dating sites compete for attention from potential matches. Member profiles provide an opportunity for candidates to present information about themselves that their counterparts use to assess compatibility and desirability. In this paper, we explore how text-based similarities among users of a dating site impact their success in attracting attention. The principle of homophily predicts that to be successful, a user should be perceived as similar to the person they would prefer to date. Conversely, theories of distinctiveness suggest that standing out from the crowd should be beneficial. Using profiles, we explore how the text similarity between a user, the opposite-sex member they are targeting, and their same-sex competitors impacts the likelihood that a sender of a message receives a response conditional on initiating contact. We find that the probability of receiving a response is maximized when the user has high text similarity to the person they message, but low text similarity to the competitors that are also seeking the same individual's attention. This suggests a balance between homophily and distinctiveness theory.

Introduction

There is little doubt that the web has become a locus of 21st century love. A vast number of romantic partnerships, perhaps now the majority, originate online. In this paper, we study the reply prediction problem on a dating site. We ask, what is the likelihood that a female user will respond to a message sent to her by a specific male candidate?

We examine data from a popular dating site in the United States. Unlike most previous research on dating sites (for exceptions, see (Fiore et al. 2010; Xia et al. 2014)), we were also able to gather users' self-created, free-text profiles.

Online dating sites are forums in which individuals attempt to manipulate the impressions that others form of them through strategic presentations of self. Verbal presentations of self are a primary means for accomplishing expressions of individuality. Here, we examine a variant of Brewer's theory of optimal distinctiveness, which posits that individuals experience tension between a desire for distinctiveness from members of an in-group, versus the need to assimilate to the group to preserve his or her inclusion (Brewer 1991). To our knowledge, this is the first test of a theory of the returns to distinctiveness in self-presentation in a online dating site.

We make two, unique contributions. First, we show that profile text similarity between a female and a male predicts her reply probability net of extensive measures of sociodemographic homophily, beauty, and other behavioral characteristics. More importantly, because we observe all searches, views, and communications, we have de facto measures of the choice set or competition experienced by users. For each female, for example, we know all males that have contacted her or that she has viewed on the site. This enables us to measure text similarities between any given male that messages a female, and all other men that are competing for her response. We assess whether it is advantageous to a male to be distinct from his competitors and we find that it is.

This set of results sheds light on open sociological questions regarding how strategic choices about self-presentation impact whether potential romantic partners become interested in one another. The results may have implications for the design of online dating platforms, which often do not consider the interplay between self-presentation, similarities among users, and the recommendation engines that suggest potential romantic partners to site users.

Related Work

To date, much of the work on online dating behavior has explored a phenomenon that is extremely well-documented in the sociological literature: across virtually every human choice network, there is evidence of homophily: people select exchange partners who are similar to them (McPherson, Smith-Lovin, and Cook 2001; Blossfeld 2009).

In analyses of dating sites, researchers in sociology, economics, and computer science have found that homogamy prevails on almost every dimension that has been studied. Site users match on geographic propinquity, race, education, income, marital histories, desire for children etc. (Fiore and Donath 2005; Hitsch, Hortaçsu, and Ariely 2010; Anderson et al. 2014; Xia et al. 2014; Lewis 2016). However, comparatively little work has investigated factors that influence the reciprocation of an attempted contact, which is the topic of our research. A notable exception is (Xia et al. 2014). The authors evaluated a series of machine learning models for predicting the likelihood of a response to a sent message using both dyadic attributes of user pairs and

standard features of the dating site network.

Following the few studies to date, we consider message response probability as a function of standard measures of sociodemographic homophily. We then extend this work to consider the effects of free-text profiles on the likelihood that a female replies to a message sent to her by a male suitor.

In addition to semantic overlap between two users, we consider the role of competition and distinctiveness as predictors of appeal. When estimating the probability that a female replies to a male, we ask whether the male fares better if he is semantically distinct from the other options that the female considers. Perhaps the most influential social psychology approach to self or social perception of differentiation is Brewer's theory of Optimal Distinctiveness. Brewer argues that conformity and differentiation are contrasting but basic human needs, which exist in opposition to one another. Specifically, on one hand, individuals feel the need to assimilate into a large, supportive group; on the other, they also experience a desire for self-enhancement, which is satisfied via membership in distinctive groups (Brewer 1991).

We argue that on a dating site, attention is more likely to flow to individuals who are somewhat differentiated from the other members of a targeted user's choice set. Conditional on appearing in a targeted user's message stream, we hypothesize that a suitor's prospects are optimized under two conditions: when the suitor is textually proximate to the target, but textually distinct from the competitors who enter the target's consideration set. In other words, we posit that the ideal strategy is to optimize cross-gender similarity, while exhibiting same-gender differentiation.

Results

Data

Our data contain 3 months (9/2013 – 11/2013) of annoymized user activity on a popular dating site in the U.S. They contain the profiles and clickstreams of 410,000 active users in ten distinct metropolitan areas. During this period, users authored 25 million messages, generated 286 million clicks on the site, and rated other users' profiles 864 million times. For each user, we have the free text content and demographic data from the user profile as well as behavioral data including complete, time-stamped browsing and rating histories. For each message sent, we know the sender, receiver and the date and time when the message was sent.

We are interested in understanding factors that affect the likelihood that a user receives a reply when they initiate contact with another user. When a user u sends a message to user v for the first time, we say that *u initiated contact* with v [1].

Based on the self-reported profile information of active users, about 55% of the users are male and 94% are heterosexual. The vast majority of messages sent during the period (93%) involve male-female dyads. Additionally, males account for 62% of all messages sent and they initiated 86% of all communication dyads. Given this over-representation of activity by males, we focus our study on male initiations and whether they are reciprocated by female users.

[1]To address right- and left-censoring in the data, we do not consider initiation messages that occurred during the first or last week of the data.

Defining Competition

When a male attempts contact with a female user, he is competing for that female's attention with other, interested males. We define three different sets of competitors based on the activity of the other males and the female. For each female user v, let I_v be the set of male users who initiated contact with v. For $u \in I_v$, let $t_{u,v}$ be the time when u initiated contact with v.

Market level Competition. We define the *market level competition* of u with respect to v, $MLC_{u,v}$, as the set of males who have ever initiated contact with v. That is $MLC_{u,v} = \{w : w \in I_v, w \neq u\}$.

We assume that u is competing with w even if w initiated contact with v after u did. The market level competition captures the type of users who tend to be interested in a particular female but not the current choices of the female.

Female Choice Competition. The *female choice competition* of u with respect to v, $FCC_{u,v}$, is defined as the set of males who initiated contact with v *before* u. That is, $FCC_{u,v} = \{w : w \in I_v, t_{w,v} < t_{u,v}\}$. This competition set captures the set of males that a female is aware of at the time an initiation occurs.

Profile View Competition. The prior competition definitions only consider the actions of the male users, but not the actions of the female user. We define a third competition set based on the user profiles the female views. For each female user v, let PV_v be the set of male users whose profile has been viewed by v. For $w \in PV_v$, let $t^{PV}_{v,w}$ be the time when v visited the profile of w. The *Profile View Competition* of u with respect to v, $PVC_{u,v}$, is defined as the set of males whose profile v has viewed before the u initiated contact with v. That is, $PVC_{u,v} = \{w : w \in PV_v, t^{PV}_{v,w} < t_{u,v}\}$. This competition set captures the set of males the female is potentially interested in based on her own activity on the site.

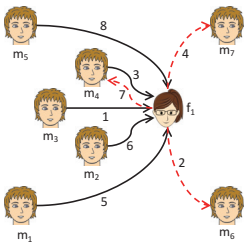


Figure 1: Illustration of three different types of competition. The black edges (solid) indicate contact with female f_1. The red edges (dashed) indicate profile visits by f_1. The label of the edges indicate the time the initiation occurred. In this example, $MLC_{m_1,f_1} = \{m_2, m_3, m_4, m_5\}$, $FCC_{m_1,f_1} = \{m_3, m_4\}$, and $PVC_{m_1,f1} = \{m_6, m_7\}$.

Measuring Profile Text Similarity

User demographic characteristics on the site we study are self-reported statements about the user's gender, sexual orientation, income, age, education level, ethnicity, geographic

location, body type, and height. To enter this information, users select from a predetermined set of options. Additionally, profiles include the opportunity to write free-text responses to multiple questions. These questions allow users to describe, in their own words, what they think is important for potential romantic partners to know about them. Responses to these questions are the means by which users describe themselves to other members of the dating site.

We use the text from these answers to measure similarity in self presentation between pairs of users. To protect user privacy we only used pseudonymized historical data. In addition, we treated each free text profile as a bag of words by removing stop words and reordering the remaining text instead of using the full free-text profile.

To measure text similarity, we converted each user's bag of words into a tf-idf vector (Jurafsky and James 2008). A keyword has a large weight insofar as it is prevalent in a focal user's response and it is infrequently used by others. Then, for each pair of site users (u,v), we computed the cosine similarity, $S_{u,v}$, between the tf-idf vectors.

The Effect of Distinctiveness on Receiving a Reply

We now explore the effect of distinctiveness in text on the likelihood of that a male user receives a reply when he initiates contact with a female. We consider two types of text similarity – *dyad text similarity* and *competition text similarity*. Dyad text similarity is the similarity between the male and the female he initiates contact with. Competition-text-similarity is the average similarity between the male who initiates with contact the female and his competition. Given a male u who initiates contact with a female v, we let $S^{MLC}_{u,v}$, $S^{FCC}_{u,v}$, $S^{PVC}_{u,v}$ be the average similarity with users in the three competition sets $MLC_{u,v}$, $FCC_{u,v}$, and $PVC_{u,v}$ respectively.

We begin by looking at the bi-variate relationship between the probability of receiving a reply and the text similarity within the dyad and between the male and his competitors across the competition networks. Figure 2(a) shows that as dyadic text similarity increases, the likelihood of a replies increases. This suggests that language-based homophily plays a role in females' decisions on who merits their attention.

Figures 2(b)-2(d) show the relationship between a male's text similarity to his competition, and the probability females reply to messages. There is a clear drop in the response probability as the male sender becomes more similar to his peers for market level and female choice sets. In other words, being distinct from other males who initiate contact with the female boosts the response probability.

The effect is less clear for text similarity in the profile view competition set. Figure 2(d) shows an initial drop in the probability of replying as similarity increases, but it is followed by an increase in probability when similarity is high. This suggests that when a female compares a focal male's profile to those of others she has browsed, he has better odds if he is either unique or prototypical.

To further test the effect that profile distinctiveness has on the probability of receiving a reply, we estimate linear probability models of the likelihood that a female responds to an initial message. Since we know from prior literature that sociodemographic homophily has a large effect on response probabilities, we include multiple, dyad-level control variables. The regression, without control variables, has the following form:

$$R_{u,v} = \beta_0 + \beta_1 S^{Comp}_{u,v} + \beta_2 S^{dyad}_{u,v} + \varepsilon_{u,v}, \quad (1)$$

where $R_{u,v}$ is 1 if the message from male u to female v received a response and 0 otherwise, $S^{Comp}_{u,v}$ is the competition similarity, and $S^{dyad}_{u,v}$ is the dyad similarity.

Control Variables. We construct a variety of measures to assess the similarity between users across the different features available on their profiles. These include both numeric features, such as person-age, and categorical features, such as ethnicity, that allow one or multiple responses.

One challenge in using the categorical options is that the choices given to users are difficult to compare. For example, for the feature *body type*, users must select one of the following options: *Athletic, Average, Fit, Little Extra, Thin, Overweight, Unknown, Skinny, Curvy, Full Figured, Won't Say, Used Up, Jacked*. It is unclear how to measure similarity or distance between the different options.

To address this issue, we use our data to measure the similarity between the possible values of different features. We start with the assumption that if two values v_1 and v_2 are similar, then users who prefer people with value v_1 will also prefer people with value v_2. Hence, we define the similarity between two values to be the fraction of females who message users with each one of the two values while controlling for what would be expected by random chance.

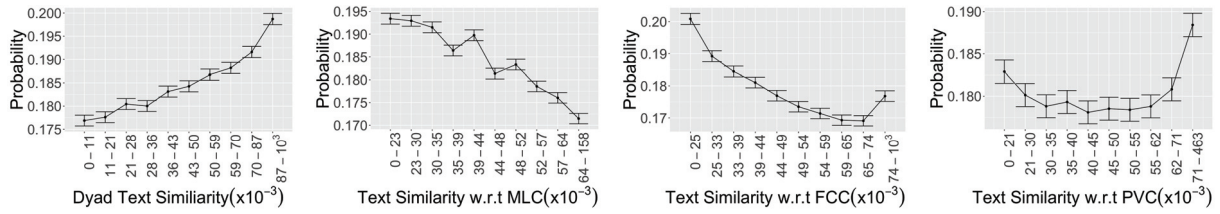
Using this approach, we measure the similarity between values for the following single-valued profile features: *body type*, *drug use*, *cigarette use*, and *ethnicity*.

For the features *language* and *type of relationship sought* that allow users to select multiple, categorical options, we use Jaccard similarity[2]. For numerical profile features, we measure distance by taking the difference between two responses. These features are *age* and *height*. Users also specify a preferred age interval, and we measure similarity by taking the overlap of the interval. Finally, users on the site are able to rate others on a 1-to-5 star scale. Guided by prior research, we use the average star rating of users as a proxy for their attractiveness (Fiore et al. 2008).

For each male u who initiated contact with a female v, we include the following control variables in our model: (i) the similarity or distance measure between each of u's and v's profile features (12 variables), (ii) an indicator variable of whether u and v live in the same city, (iii) dyad text similarity ($S_{u,v}$), (iv) the difference in attractiveness between u and the average attractiveness in the three competition sets $MLC_{u,v}$, $FCC_{u,v}$, and $PVCu,v$, (v) the age of v, and (vi) percentage of initiations v replied to. All variables that involve taking a difference are defined as the value for the male u minus the value of the female v or the value of the average competitor.

Table 1 shows the results from the linear probability model for each of the competition sets. We highlight the following observations. First, among the dyad similarity and

[2]The Jaccard similarity of two sets A and B is defined as $\frac{A \cap B}{A \cup B}$



(a) Dyad (b) Market Level (c) Female Choice (d) Profile View

Figure 2: Variation of the probability of a response with text similarity

	MLC	FCC	PVC
Intercept	-4.04e-15	1.447e-15	7.635e-15
dyad age difference	-0.0379***	-0.0406***	-0.0416***
dyad height difference	0.0126***	0.0130***	0.0177***
dyad language similarity	0.0065***	0.0078***	0.0060***
dyad sought relationship type similarity	0.0009	-0.0001	-0.0001
dyad physical distance indicator	0.0070***	0.0074***	0.0098***
dyad preferred age interval overlap	0.0179***	0.0174***	0.0179***
dyad body-type similarity	-0.0028***	-0.0035***	-0.0037***
dyad drugs similarity	0.0041***	0.0032***	0.0027***
dyad ethnicities similarity	0.0192***	0.0185***	0.0227***
dyad smokes similarity	0.0076***	0.0075***	0.0062***
dyad attractiveness difference	0.0640***	0.0339***	0.0599***
dyad text similarity	0.0243***	0.0169***	0.0108***
MLC attractiveness	0.1162***	-	-
FCC attractiveness	-	0.1354***	-
PVC attractiveness	-	-	0.1118***
MLC text similarity	-0.0289***	-	-
FCC text similarity	-	-0.0222***	-
PVC text similarity	-	-	-0.0074***
female message response rate	0.3720***	0.3683***	0.3509***
age of female	-0.0319***	-0.0264***	-0.0294***
* p <0.05 ** p <0.01 *** p <0.005			

Table 1: OLS regression coefficients for competition sets

difference variables, all except for height, body type, and attractiveness support homophily – the female user is more likely to respond when the male is similar to her. It is natural to expect that when the male has a more desirable features in terms of height, body type and attractiveness, the female is more likely to respond despite the male being different from her. Second, the male is more likely to receive a response when he is more attractive than his average competitor. Finally, the male is more likely to receive a reply when he is different from his average competitor.

These results show that there is a robust relationship between a male's distinctiveness and the likelihood of capturing the interest of the female, net of several control variables that account for homophily, the rate at which the female responds to first time contacts, and attractiveness. Importantly, these effects are similar for different types of competition sets, which further highlight the robustness of the finding.

Conclusion

Online dating sites represent a vast opportunity for large-scale, quantitative studies of the ways in which social actors interact to create and shape social perceptions of selves. Members of these sites have many degrees of freedom in crafting online identities with the goal of appealing to other members of the community.

We find that male-female textual similarity increases the likelihood of a match as evidenced by a reciprocated communication link. Conversely, we find that there is a dividend to differentiation from same-sex competition: males who describe themselves in language that is distinct from the other men they compete with are more likely to be rewarded with a response. This suggests that the optimal strategy is a balancing act: exhibit common interests with the opposite sex, while standing out from one's own.

References

Anderson, A.; Goel, S.; Huber, G.; Malhotra, N.; and Watts, D. J. 2014. Political ideology and racial preferences in online dating. *Sociological Science* 1:28–40.

Blossfeld, H.-P. 2009. Educational assortative marriage in comparative perspective. *Annual review of sociology* 513–530.

Brewer, M. B. 1991. The social self: On being the same and different at the same time. *Personality and social psychology bulletin* 17(5):475–482.

Fiore, A. T., and Donath, J. S. 2005. Homophily in online dating: when do you like someone like yourself? In *CHI'05 Extended Abstracts on Human Factors in Computing Systems*, 1371–1374. ACM.

Fiore, A. T.; Taylor, L. S.; Mendelsohn, G. A.; and Hearst, M. 2008. Assessing attractiveness in online dating profiles. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 797–806. ACM.

Fiore, A. T.; Taylor, L. S.; Zhong, X.; Mendelsohn, G. A.; and Cheshire, C. 2010. Who's right and who writes: People, profiles, contacts, and replies in online dating. In *hicss*, 1–10.

Hitsch, G. J.; Hortaçsu, A.; and Ariely, D. 2010. What makes you click?mate preferences in online dating. *Quantitative marketing and Economics* 8(4):393–427.

Jurafsky, D., and James, H. 2008. Speech and language processing an introduction to natural language processing, computational linguistics, and speech.

Lewis, K. 2016. Preferences in the early stages of mate choice. *Social Forces* sow036.

McPherson, M.; Smith-Lovin, L.; and Cook, J. M. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology* 415–444.

Xia, P.; Jiang, H.; Wang, X.; Chen, C. X.; and Liu, B. 2014. Predicting user replying behavior on a large online dating site. In *ICWSM*. Citeseer.