# Homophily in Online Dating: When Do You Like Someone Like Yourself?

**Andrew T. Fiore and Judith S. Donath**
MIT Media Laboratory
20 Ames St., Cambridge, Mass., USA
{fiore, judith}@media.mit.edu

**ABSTRACT**
Psychologists have found that actual and perceived similarity between potential romantic partners in demographics, attitudes, values, and attractiveness correlate positively with attraction and, later, relationship satisfaction. Online dating systems provide a new way for users to identify and communicate with potential partners, but the information they provide differs dramatically from what a person might glean from face-to-face interaction. An analysis of dyadic interactions of approximately 65,000 heterosexual users of an online dating system in the U.S. showed that, despite these differences, users of the system sought people like them much more often than chance would predict, just as in the offline world. The users' preferences were most strongly same-seeking for attributes related to the life course, like marital history and whether one wants children, but they also demonstrated significant homophily in self-reported physical build, physical attractiveness, and smoking habits.

**Author Keywords**
Online personals, attraction, computer-mediated communication, online dating, relationships

**ACM Classification Keywords**
H5.3. Group and Organization Interfaces; Asynchronous interaction; Web-based interaction.

## INTRODUCTION

Online personal advertisements — lengthier, more detailed descendants of newspaper personal ads — have grown rapidly in recent years. In August 2003, personals Web sites in the United States drew 40 million unique visitors — half the number of single adults in the U.S. (Mulrine 2003).

These online personal ads have shed their stereotype as matchmakers for the awkward and now claim a prominent role in the social lives of millions of users. Researchers have studied online friendships and romantic relationships from psychological and sociological perspectives (Lea & Spears 1995, Walther 1996, McKenna et al. 2002), and they have examined the personals ads that appear in print publications (Bolig et al. 1984, Ahuvia & Adelman 1992). This paper describes a quantitative examination of the characteristics for which online dating users seek others like them.


*CHI 2005*, April 2–7, 2005, Portland, Oregon, USA.
ACM 1-59593-002-7/05/0004.

## NATURE OF ONLINE PERSONALS DATA

We analyzed data from one online dating system in particular. Through an agreement brokered by the Media Laboratory with an online dating Web site (the "Site"), we obtained access to a snapshot of activity on the Site over an eight-month period, from June 2002 through February 2003. The data included users' personal profile information, their self-reported preferences for a mate, and their communications via the site's private message system with other users. Anonymous ID numbers distinguished unique users.

Table 1 indicates which profile characteristics users could specify about themselves and about the partners they would like to meet.

Data about private messages exchanged by the users included the sender, recipient, subject, text, date and time of delivery, and whether the recipient had read the message.

## USER DEMOGRAPHICS

The Site had 221,800 members as of February 2003, the end of the eight-month period covered by our snapshot of the Site's data. Of these, 69.4 percent (153,942 users) had fully completed their profiles. A slightly different subset, 25.9 percent of the total (57,362 users), was active during the eight-month study period — that is, they sent or received at least one message. This active subset was used for analyses of messaging behavior, but for analyses involving profile characteristics, it was limited to the 23.8 percent of users (52,857) who were active and had complete profiles.

Although the Site has a national base of users, they are distributed differently from the U.S. population on a state-by-state basis. Heavy use appears in upstate New York, the Southeast, the Midwest and Great Lakes regions, and certain secondary urban areas in the West, such as Sacramento, Calif.

| Attribute | Type |
|---|---|
| Online handle | Free |
| Gender | Cat. |
| Age | Con. |
| Height | Con. |
| Location (city, state, postal code) | Cat./Free |
| Physical build | Cat. |
| Drinking habits | Cat. |
| Smoking habits | Cat. |
| Educational level | Cat. |
| How many children user has | Buck. |
| How many children user wants | Buck. |
| Marital status | Cat. |
| Pets owned | Cat. |
| Pets preferred | Cat. |
| Self-rated physical attractiveness | Buck. |
| Race | Cat. |
| Type of relationship sought | Cat. |
| Religion | Cat. |
| Importance of age in a partner | Buck. |
| Importance of height in a partner | Buck. |
| Textual self-description | Free |

[Type: *Free* text, *Cat*egorical, *Con*tinuous, *Buck*eted]

**Table 1.** Profile attributes specified by users about themselves and about their preferences in a partner

The overall user population on the Site included more men (62.8 percent) than women (37.2 percent), but the active subset analyzed in this work was 55 percent female.

The Site targets heterosexual users; although it allows users to specify same-sex preferences (e.g., "male seeking male"), less than one percent of users did so, and many of these appeared to be data entry mistakes or confusion about the interface. Because homosexual users were so few, their behavior would be inadequate to draw conclusions about gay users' behavior in online dating environments; thus, these users were excluded from the analysis.

Within the active subset of users, the median age was 34, but the male population was slightly older (median 36 years, compared to 33 years for women). Most users were Caucasian (83.7 percent); African-Americans and Hispanics each composed approximately two percent of the user population. Nearly 10 percent chose not to give their race.

For additional demographic descriptors of this data set, including religion, marital status, number of children, educational level, smoking habits, drinking habits, physical build, and physical attractiveness, consult Fiore (2004).

## MESSAGES AND CONVERSATIONS

During the eight-month period from June 2002 to February 2003, 29,687 users sent 236,930 messages to 51,348 users. In total, these messages constituted 110,722 exchanges of one or more messages between unique pairs of users (a *conversation*). However, most of these exchanges were something less than dyadic: 78.2 percent (86,597) of conversations consisted of unreciprocated single messages.

Messages were received in a more even distribution than they were sent; that is, fewer members sent messages than received them. Users sent and received a mean of 1.50 messages (median = 0.0) in the eight-month study period. The means are the same because the same bounded population sent and received the messages. However, the standard deviation for number of messages sent was 7.45, as compared to 4.90 for number of messages received, indicating that messages were distributed more evenly across the set of recipients than they were across the set of senders. In total, 29,687 users sent 236,930 messages to 51,348 users.

Of exchanges between a man and a woman, men initiated the majority of conversations (73.3 percent vs. 26.7 percent); however, their initiations were 17.9 percent less likely to be reciprocated than those begun by women (20.6 percent reciprocated vs. 25.1 percent for female-initiated; $t = -15.465$; d.f. = 50,150; $p < 0.001$).

Users of both sexes had contact with a median of 2.0 distinct others. The distribution was wider, though, for men than for women (mean = 5.3, s.d. = 11.8 for men; mean = 4.2, s.d. = 5.9 for women). Men participated in more communications on average than women, but we would expect this because the active subset of users contains more women than men, so the contacts are spread across a larger number of women.

Overall, the number of dyadic ties per person followed the familiar "power law" distribution, with many users with few ties and exponentially fewer with many ties. The mean number of ties per person was 5.0 (s.d. = 9.04).

As expected from the above finding that men begin most conversations, men on average initiated more contacts than they received (mean = 3.3, median = 1.0, s.d. = 7.1 initiated vs. mean = 1.9, median = 1.0, s.d. = 2.8 received). Women, on the other hand, initiated fewer contacts than they received (mean = 1.5, median = 0.0, s.d. = 3.4 initiated vs. mean = 2.7, median = 1.0, s.d. = 3.5 received).

## SEEKING HOMOPHILY: WHICH DIMENSIONS?

Some characteristics are more bounding than others — that is, users are more likely to seek someone like themselves on that dimension. For example, smokers might want to find other smokers more so than people with blue eyes want to find other people with blue eyes. We would say that smoking is more strongly bounding than eye color because people with a given smoking status are less likely to cross the boundary to choose someone with a different smoking status than someone with brown eyes would be to choose a partner with blue eyes.

To determine the bounding strength of categorical and bucketed descriptors in the data set, we compared the

percentage of contacts between two users who shared the same value for a characteristic (e.g., "athletic" for the characteristic "physical build") with the percentage of contacts we would *expect* to share the value if one male user and one female user from the active user population were paired randomly.

**Analytic Approach**

On the Site, 32.6 percent of male users and 9.2 percent of female users report their build as "athletic." If users were contacting each other randomly but in heterosexual pairs, we would expect 0.326 * 0.092, or 3.0 percent, of contacts to involve two users of athletic build. However, if users of athletic build sought other such users more often, the percentage of contacts involving two of these users would exceed 3.0 percent; if these users avoided each other, the percentage would be lower.

By summing the probability of sameness across all possible values of a characteristic, we find an overall probability that a random pair of one male and one female user will share the same value for that characteristic. These overall probabilities are listed in Table 2 as *Expected percent same*. The expected sameness for a characteristic varies with the number of values possible for that characteristic and how evenly users are distributed among the values. Expected sameness is higher when the number of values is low, as with Physical Appearance ("Very attractive," "Attractive," "Average," "Prefer not to answer"), and when many users have picked the same value for a characteristic, as with Race (83.7 percent reported "Caucasian").

Having calculated the expected sameness, we computed the actual percentage of dyads with the same value for each categorical characteristic both for all pairwise exchanges and separately for the subset of reciprocated exchanges. The absolute value of the difference between the actual percentage of sameness and the expected percentage of sameness indicates how much users were deliberately seeking someone with the same value as themselves.

An actual sameness percentage close to its expected sameness percentage indicates that users who share a value for that characteristic did not communicate more often than we would expect by chance if users were contacting each other randomly. On the other hand, a large difference between actual and expected sameness percentages would indicate that users who share a value for a characteristic communicated more often than we would expect by chance.

Because we expect statistically a varying likelihood of sameness for various characteristics, the absolute difference in expected and actual percentages does not facilitate comparisons between different characteristics, which have different expected percentages. Instead, we calculate the proportion of the actual to the expected percentage sameness for each characteristic. Table 2 shows these values in parentheses following the actual percentages for all contacts

| Characteristic | Expected percent same ($x$) | Actual percent same (all contacts, $a_1$) | Actual percent same (recip. con. only, $a_2$) | t ($a_2 \neq x$) |
|---|---|---|---|---|
| Marital status | 31.6 | 51.7 (1.64x) | 56.0 (1.77x) | 76.001† |
| Wants children | 25.1 | 38.7 (1.54x) | 40.5 (1.61x) | 48.553† |
| Num. of children | 27.8 | 38.7 (1.39x) | 38.6 (1.39x) | 34.352† |
| Physical build | 19.2 | 24.5 (1.28x) | 25.6 (1.33x) | 22.435† |
| Smoking | 40.5 | 50.6 (1.25x) | 54.0 (1.33x) | 41.979† |
| Phys. appearance | 37.6 | 46.1 (1.23x) | 49.2 (1.31x) | 35.886† |
| Educational level | 23.6 | 28.0 (1.19x) | 29.3 (1.24x) | 19.360† |
| Religion | 42.4 | 49.7 (1.17x) | 52.6 (1.24x) | 31.589† |
| Race | 71.1 | 81.2 (1.14x) | 85.9 (1.21x) | 65.808† |
| Drinking habits | 61.2 | 68.7 (1.12x) | 73.4 (1.20x) | 42.692† |
| Pet preferences | 34.7 | 38.5 (1.11x) | 39.9 (1.15x) | 16.425‡ |
| Pets owned | 21.8 | 23.6 (1.08x) | 24.0 (1.10x) | 8.038‡ |

† d.f. = 23,940; p < 0.001  ‡ d.f. = 23,855; p < 0.001

**Table 2.** Bounding strength of categorical characteristics. *Expected percent same* indicates the statistically expected percentage of dyadic pairs who share the same value for the listed characteristic. The expected probability is based on random selection from the male and female population distributions for the characteristic. *Actual percent same* indicates the empirical percentage of dyadic pairs who shared the same value for the listed characteristic, across all contacts and just the reciprocated subset, in which the initial recipient replied.

and for reciprocated contacts. The characteristics are listed in descending order of this proportion, which shows the relative bounding strength of each.

### Findings

Users opted for sameness more often than chance would predict in all the characteristics examined in this section. This concurs with the overwhelming evidence gathered by relationship researchers (cf. Brehm et al. 2002, Fisher 1992) that actual and perceived similarity in demographics, attitudes, values, and attractiveness correlate with attraction (and, later, relationship satisfaction). However, users demonstrate this homophily to differing degrees for different characteristics.

Dyads were much more likely than chance to choose the same value for characteristics relating to the life course. Values for marital status and wanting children were the same in dyads 64 percent and 54 percent more often, respectively, than would occur with random pairings. The number of children users already have was the same in dyads 39 percent more often than chance. These were the three most strongly bounding characteristics.

Physical build was the same among dyads 28 percent more often than chance would predict. This finding rests on similarity-seeking among a few possible values for build, such as "average" and "athletic," that encompass both genders; many of the other possibilities, like "petite" and "body-builder," are strongly gendered and thus very unlikely to be the same in a heterosexual dyad. Physical appearance, a self-reported rating of attractiveness, was the same among dyads 23 percent more often than chance. Among lifestyle choices, including smoking habits, drinking habits, and pet preferences, only smoking was the same in dyads more than 20 percent more often than chance would predict. Most dyads (68.7 percent) were the same in drinking habits, but this is because 75.6 percent of men and 77.9 percent of women identified themselves as "Social/occasional" drinkers. Thus, the expected probability of sameness was also high for this characteristic, rendering the high actual similarity unremarkable.

Pets, both general preferences regarding them and specific pets already owned, proved the least bounding of any characteristics. Users picked others who shared their preferences only about 10 percent more often than chance would predict. Homogeneity on these characteristics did not matter to users nearly as much as other characteristics.

Religion was the same in dyads 17 percent more often than chance. More than half of active users of the Site identified themselves as Christian, and about a third chose "Prefer not to answer," a very high percentage compared to other characteristics. Given the distribution of religions among users who did answer, we might reasonably presume that a large number of "Prefer not to answer" respondents are in fact Christians, even if we allow that non-Christians might be more likely to choose "Prefer not to answer." If this is the case, the bounding strength of religion might appear lower than it is because of users' reluctance to specify their religion. It might also be true that having similar religiosity is more important than sharing a specific religion (cf. Williams & Lawler 2003).

The overwhelming majority of dyads (81.2 percent) shared the same race, but, as with drinking habits, this high rate of similarity is only moderately better than chance (14 percent). Because 83.7 percent of users were Caucasian, the rate of similarity expected by chance was also high, 71.1 percent.

Characteristics were slightly more bounding among the subset of reciprocated contacts, but the difference was small and roughly equal across characteristics. Although the difference is small, it suggests that users were slightly more likely to respond to an initiation from a more similar other.


### ACKNOWLEDGMENTS

We would like to thank Dan Ariely and Jeana Frost for their insights in this work.


### REFERENCES


1. Ahuvia, A.C., & M.B. Adelman. "Formal Intermediaries in the Marriage Market: A Typology and Review." In *J. of Marriage and the Family* 54 (May 1992): 452-463.
2. Bolig, R., Stein, P.J., & P.C. McKenry. "The Self-Advertisement Approach to Dating: Male-Female Differences." In *Family Relations* 33 (1984): 587-592.
3. Brehm, Miller, Perlman, and Campbell (2002). *Intimate Relationships* 3e. New York: McGraw-Hill.
4. Fiore, A.T. *Romantic Regressions: An Analysis of Behavior in Online Dating Systems*. S.M. thesis, Massachusetts Institute of Technology, Sept. 2004.
5. Fiore, A.T., & J.S. Donath. "Online Personals: An Overview." Short paper, *CHI 2004*, Vienna, Austria.
6. Fisher, H. Anatomy of Love: A natural history of mating, marriage, and why we stray. New York: Fawcett-Columbine, 1992.
7. Lea, M., & Spears, R. "Love at first byte? Building personal relationships over computer networks." In J. T. Wood & S. Duck (Eds.), *Under-studied relationships: Off the beaten track* (pp. 197-233). Thousand Oaks, CA: Sage, 1995.
8. Mulrine, A. "Love.com: For better or for worse, the Internet is radically changing the dating scene in America." *U.S. News & World Report*, Sept. 29, 2003.
9. Walther, J. B. "Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction." In *Comm. Research* 23, 3-43, 1996.
10. Williams, L.M. & M.G. Lawler. "Marital Satisfaction and Religious Heterogamy: A Comparison of Interchurch and Same-Church Individuals." In *J. of Family Issues* 24 (8), 2003: 1070-1092