sociological science

Supplement to:

Bruch, Elizabeth E., and M. E. J. Newman. 2019. "Structure of Online Dating Markets in U.S. Cities." Sociological Science 6: 219-234.

Structure of online dating markets: Supplementary Information

Elizabeth E. Bruch and M. E. J. Newman

1 Data

Our data come from a popular, free online dating site. New users of the site begin by creating a profile, which includes various socio-demographic information, and they can also answer a set of open-ended essay questions that ask them to describe who they are and what they are looking for. The only information a user is required to give is their login handle, age, sexual orientation, relationship status, and a 5-digit ZIP code identifying their location. After creating a profile, users can then view the profiles of others, as well as send and receive messages. Unlike other dating sites, that are largely driven by a matching algorithm, our site allows users to pursue mates relatively freely according to their own preferences.

1.1 Metropolitan Areas

Our city-level results are based on data from four metropolitan areas—New York City, Boston, Chicago, and Seattle. In the case of Boston, Chicago, and Seattle, we find a good choice of boundaries to be the standard Core Based Statistical Areas (CBSAs) established by the Office of Management and Budget.[1] For New York City, however, the data clearly indicate multiple geographic dating markets within the larger metro area. Instead, therefore, we choose a narrower

[1] A CBSA is defined to be an urban center of at least 10 000 people plus adjacent areas that are socioeconomically tied to the urban center by commuting.

	New York		Boston		Chicago		Seattle	
	Men	Women	Men	Women	Men	Women	Men	Women
Total number of users	44 009	50 618	9 113	9 355	28 635	23 236	12 721	9 248
Ethnicity (%)								
Asian	8	11	4	6	3	4	7	9
Black	9	9	6	6	7	9	4	3
Hispanic	10	8	3	3	8	7	3	3
White	73	73	87	85	81	80	87	85
College degree (%)	92	96	70	80	63	71	64	68
Children at home (%)	5	6	7	10	7	10	15	17
Mean age	31.6	31.5	30.4	30.3	31.4	32	32.7	33.1
Mean messages sent	23.3	9.4	14.6	6.3	19	10.2	12.4	7.8
Replies received (%)	15	34	17	37	18	40	20	45

Table S1: User attributes for four metropolitan areas. Table reproduced from Ref. (*1*)

set of geographic boundaries for New York, the five boroughs of Manhattan, Brooklyn, Queens, the Bronx, and Staten Island.

1.2 Summary statistics

Table S1 provides summary statistics of users in each of the four cities, broken out by gender. As discussed in the paper, the cities vary in the ratio of men to women on the web site, New York having the largest fraction of women, followed by Boston, Chicago, and Seattle, in that order. Recall from Figure 2B that we found the older submarkets to be more female-heavy, while the younger submarkets tended to be male-heavy. Figure S1, which shows the age distribution of men and women in each city, suggests that this is not merely a result of age-specific sex ratios in the overall user population. We observe that New York, for instance, has a surplus of women, which is most pronounced among younger users in their mid twenties, yet the submarkets for younger users still have significantly more men than women. (The remaining cities all have an overall surplus of men, which is most pronounced in the later 20s and early 30s.) These observations suggest that the submarket sex ratios observed in Figure 2B are driven by users'

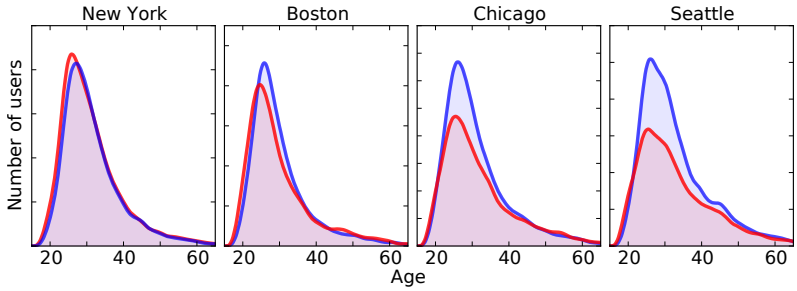

Figure S1: Age distribution of men and women in each city. Boston, Chicgao, and Seattle all have surpluses of men, the surplus being most pronounced for people around 28 years of age. New York city has a surplus of women, which is most pronounced among people in their mid-twenties. Note that because the total number of users varies across cities, the scale of the y-axis differs across the four panels. Figure reproduced from Ref. (*1*)

mate seeking behavior, and not broader population demographics.

In addition to the sex ratios, Table S1 also shows that cities differ in their overall market size and composition. New York City is the largest market, followed by Chicago, Seattle, and Boston. We also observe some variation in the average number of initial contacts made by men and women in each city, as well as their reply rates. Consistent with other work (*2–4*), we see that men send more messages than women. However, men have a lower chance than women of receiving replies to their messages.

2 Network analysis

As described in the paper, the starting point for our results is community structure analysis of networks of reciprocated messaging between pairs of individuals. Our city-level analyses are restricted to the largest connected component of the network for each city, although in practice this has little effect since nearly everyone belongs to the largest component. In the network for New York City, for example, the largest connected component contains 99.8% of all users.

Our analysis of the full, nationwide messaging network in Fig. 1 of the main paper is based

on standard modularity maximization, as described in the Methods. The structure within our individual city networks, however, is more complicated, being partly assortative (with respect to submarket) but also partly disassortative (with respect to gender, since most messages are between a man and a woman). To correctly detect and classify this kind of mixed structure we need a more flexible detection method. The leading such method is the statistical inference method based on fitting the network to a stochastic block model (*5–8*), which is the approach we employ in this work. Specifically, we use the degree-corrected stochastic block model (*7*), which is a generative model of a random community-structured network as follows.

Let n be the number of nodes in the observed network (a number typically in the thousands or tens of thousands for the networks studied here). The degree-corrected block model allows us to create a model network of the same size by first generating n nodes, numbered from 1 to n, each of which is assigned to one of k communities or submarkets. The communities are numbered from 1 to k, and nodes are assigned to communities independently at random, with probability γ_r of being assigned to community r, where the γ_r are parameters we choose, subject to the normalization constraint

$$\sum_{r=1}^{k} \gamma_r = 1. \tag{S1}$$

When all nodes have been assigned to communities, edges are placed at random between pairs of nodes, independently but with probabilities that depend on the communities to which the nodes belong, such that when all edges have been placed the number falling between any pair of nodes i, j is Poisson distributed with mean $d_i d_j \omega_{rs}$, where r and s are, respectively, the communities to which nodes i and j belong, ω_{rs} are parameters that we choose, and d_i is the degree of node i in the observed network that we are fitting (i.e., it is the number of connections node i has to other nodes). The inclusion of d_i is what distinguishes this "degree-corrected" model from other forms of the stochastic block model. As we will see, the degree correction fixes the expected degree of every node within the model to be equal to the observed degree of the same node in

the data, allowing the model to give significantly better fits to empirical data.

This defines the "forward" process of generating a random network given the parameters γ, ω of the model. Using the model for community detection involves the inverse process of fitting the model to observed data so as to determine the values of the parameters that give the best fit. This we do by the method of maximum likelihood. Our undirected network of two-way communication between web site users is represented by an adjacency matrix A with elements $a_{ij} = 1$ if there is an edge between nodes i and j and zero otherwise. It is straightforward to show that the probability, or likelihood, of generating the observed network from the model, for given values of the parameters γ, ω, is

$$P(A|\gamma,\omega) = \sum_c P(A, c|\gamma,\omega) = \sum_c \mathrm{e}^{\mathscr{L}(c)}, \tag{S2}$$

where c denotes the complete set of community assignments $\{c_i\}$ and the log-likelihood $\mathscr{L}(c) = \log P(A, c|\gamma,\omega)$ of generating a particular set of community assignments and edges is given by

$$\mathscr{L}(c) = \sum_{ij}\bigl[a_{ij}\log\omega_{c_i,c_j} - d_i d_j \omega_{c_i,c_j}\bigr] = \sum_{ijrs}\delta_{c_i,r}\delta_{c_j,s}\bigl[a_{ij}\log\omega_{rs} - d_i d_j \omega_{rs}\bigr], \tag{S3}$$

where δ_{rs} is the Kronecker delta and we have neglected additive and multiplicative constants independent of the parameters, since they have no effect on the position of the likelihood maximum.

2.1 Expectation-maximization (EM) algorithm

To find the values of the parameters γ and ω most likely to have generated the observed network we wish to maximize equation (S2) with respect to the parameters. Direct maximization is cumbersome so we employ a standard trick from the machine learning toolkit. First, we maximize not the likelihood itself but its logarithm, $\log P(A|\gamma,\omega)$, which gives the same result since the logarithm is a monotone increasing function of its argument and hence the maximum of

the logarithm falls in the same place as the maximum of the argument. Then we apply Jensen's inequality, which says that for any set of nonnegative quantities x_i, we have

$$\log \sum_i x_i \geq \sum_i q_i \log \frac{x_i}{q_i}, \tag{S4}$$

where q_i is any properly normalized probability distribution satisfying $\sum_i q_i = 1$. The exact equality is recovered for the special choice

$$q_i = \frac{x_i}{\sum_i x_i}. \tag{S5}$$

Applying Jensen's inequality to the log of equation (S2), we find that

$$\log P(A|\gamma,\omega) = \log \sum_c \mathrm{e}^{\mathscr{L}(c)} \geq \sum_c q(c) \log \frac{\mathrm{e}^{\mathscr{L}(c)}}{q(c)} \tag{S6}$$

$$= \sum_{ijrs} q_{rs}^{ij} \left[a_{ij} \log \omega_{rs} - d_i d_j \omega_{rs} \right] - \sum_c q(c) \log q(c), \tag{S7}$$

where we have made use of equation (S3) for the log-likelihood $\mathscr{L}(c)$. Here $q(c)$ is any properly-normalized probability distribution we choose over community assignments c, and q_{rs}^{ij} is the probability within that distribution that nodes i and j belong to communities r and s respectively, thus:

$$q_{rs}^{ij} = \sum_c \delta_{c_i,r} \delta_{c_j,s}\, q(c). \tag{S8}$$

Following equation (S5), the exact equality in (S7) is established, and hence the right-hand side maximized, when we make the choice

$$q(c) = \frac{P(A,c|\gamma,\omega)}{\sum_{c'} P(A,c'|\gamma,\omega)} = \frac{\mathrm{e}^{\mathscr{L}(c)}}{\sum_{c'} \mathrm{e}^{\mathscr{L}(c')}}. \tag{S9}$$

Thus if we maximize the right-hand side of (S7) over possible choices of $q(c)$ it becomes equal to the left-hand side, and if we further maximize the left-hand size with respect to the parameters γ, ω we get the answer we are looking for—the values of γ, ω that maximize the overall likelihood. Put another way, a double maximization of the right-hand side with respect to both $q(c)$ and ω, γ will achieve our goal.

At first sight, this appears to make the problem harder: we have turned what was previously a single maximization into a double one. But in fact the double maximization usefully splits the problem into two parts that separately are both straightforward, whereas the original combined problem was difficult. Maximization with respect to $q(c)$ is achieved by making the choice (S9), as we have said. Maximization with respect to γ and ω can be achieved by simple differentiation. Note that the final sum on the right-hand side of equation (S7) does not depend on γ or ω, so it vanishes upon differentiating. Taking the derivative of the first sum with respect to γ_r and ω_{rs} while imposing the constraint (S1) then gives us

$$\gamma_r = \frac{1}{n} \sum_i q_r^i, \tag{S10}$$

and

$$\omega_{rs} = \frac{\sum_{ij} a_{ij} q_{rs}^{ij}}{\sum_i d_i q_r^i \sum_j d_j q_s^j}, \tag{S11}$$

where q_r^i is the probability within the distribution $q(c)$ that node i belongs to group r:

$$q_r^i = \sum_c \delta_{c_i,r} q(c) = \sum_s q_{rs}^{ij}, \tag{S12}$$

the second equality being true for any value of j.

The result is an expectation-maximization or EM algorithm for fitting the model to the observed network, requiring the simultaneous solution of equations (S9), (S10), and (S11), which is accomplished by simple iteration. We first choose initial values of the parameters γ and ω, for instance at random, and use them to calculate the probability distribution $q(c)$ from equation (S9). Then we use that distribution to calculate q_{rs}^{ij} and q_r^i from equations (S8) and (S12), and thence to calculate improved estimates of the parameters from equations (S10) and (S11). Then we recalculate $q(c)$ again, and repeat until convergence is reached.

The end product is a set of best-fit values of the parameters to the observed network data. In addition to this, however, and crucially for our purposes, we also calculate a converged value of

the distribution $q(c)$, which, from equation (S9), is equal to

$$q(c) = \frac{P(A, c|\gamma, \omega)}{\sum_{c'} P(A, c'|\gamma, \omega)} = \frac{P(A, c|\gamma, \omega)}{P(A|\gamma, \omega)} = P(c|A, \gamma, \omega). \tag{S13}$$

In other words, $q(c)$ is the *posterior distribution* over community assignments, the probability, given the observed data A and the best-fit parameter values, of any particular division c of the network into communities. The final step of the calculation is then to assign each node to the community for which it has the highest probability of membership, which is also equivalent to choosing the community for which q_r^i is maximized. This gives us our best division of the network into communities or submarkets.

2.2 Expected degree

A key feature of the degree-corrected block model is its ability to provide a good fit to networks with broad distributions of node degree (the degree of a node in a network being the number of connections it has to other nodes). Most empirical networks, including our messaging networks, have widely varying values of node degree and any model we fit to such networks must, at a minimum, be capable of capturing this variation.

The actual degree of a node in our model network can fluctuate from one realization of the model to another, since the model contains random elements. But the expected value of the degree of node i, for the best-fit values of the parameters γ, ω given in Eqs. (S10) and (S11), is always equal to the degree d_i of the same node in the observed network. Thus the fitted network fits the degree distribution exactly apart from fluctuations. To see this, observe that the expected degree of node i in the model is equal to the sum of the expected number of edges $d_i d_j \omega_{c_i,c_j}$ between node i and every other node $\sum_j d_i d_j \omega_{c_i,c_j}$, averaged over the distribution $q(c)$ of community assignments, thus:

$$\sum_c q(c) \sum_j d_i d_j \omega_{c_i,c_j} = \sum_c q(c) \sum_j d_i d_j \sum_{rs} \delta_{c_i,r} \delta_{c_j,s} \omega_{rs} = \sum_{jrs} q_{rs}^{ij} d_i d_j \omega_{rs}, \tag{S14}$$

where we have made use of equation (S8). Most nodes j, however, will be far from node i in a large network, so that the community assignments of i and j are essentially uncorrelated. This means that $q_{rs}^{ij} = q_r^i q_s^j$ and the expected degree becomes

$$d_i \sum_{rs} q_r^i \omega_{rs} \sum_j q_s^j d_j = d_i \sum_{rs} \frac{q_r^i \sum_{ij} a_{ij} q_{rs}^{ij}}{\sum_k d_k q_r^k} = d_i \sum_r \frac{q_r^i \sum_{ij} a_{ij} q_r^i}{\sum_k d_k q_r^k}$$
$$= d_i \sum_r \frac{q_r^i \sum_i d_i q_r^i}{\sum_k d_k q_r^k} = d_i \sum_r q_r^i = d_i, \tag{S15}$$

where we have made use of equation (S11) in the first equality, equation (S12) in the second, and the trivial observation $\sum_j a_{ij} = d_i$ in the third.

2.3 Belief propagation and the calculation of the posterior distribution

Elegant though the EM algorithm is for the community detection problem, it is not (yet) a workable method, because for all but the very smallest of networks is it not feasible to evaluate the posterior distribution $q(c)$ directly from equation (S9)—the number of possible values of c is simply too large. The number of possible divisions of n nodes into k communities is k^n, so a division of 10 000 nodes into, say, four communities would have $4^{10000} \simeq 10^{6000}$ possible divisions, which is far more than can be enumerated by even the most powerful computer. Within the statistical literature, the standard way of circumventing this problem is to approximate the distribution $q(c)$ using Markov chain Monte Carlo importance sampling, and that could be done here too. In our work, however, we use a recently-proposed alternative approach based on belief propagation (*8–10*), which is significantly more efficient for the particular problem at hand.

The belief propagation method focuses on a quantity $\mu_r^{i\to j}$, called the belief, which is equal to the (posterior) probability that node i belongs to community r if we are not told whether there is an edge between nodes i and j, i.e., if we are given the entire adjacency matrix A except for the element a_{ij}. The omission of this one matrix element is crucial to the method: it allows us to write a self-consistent set of equations for the beliefs that can be solved by numerical iteration.

For the degree-corrected block model used here, the appropriate equations have been given by Yan *et al.* (*10*):

$$\mu_r^{i\to j} = \frac{\gamma_r}{Z_{i\to j}} \exp\biggl(-\sum_k d_i d_k \sum_s \omega_{rs} q_s^k\biggr) \prod_{\substack{k(\neq j)\\ a_{ik}=1}} \omega_{rs}\mu_s^{k\to i}, \tag{S16}$$

where $Z_{i\to j}$ is a normalizing constant with value

$$Z_{i\to j} = \gamma_r \sum_r \exp\biggl(-\sum_k d_i d_k \sum_s \omega_{rs} q_s^k\biggr) \prod_{\substack{k(\neq j)\\ a_{ik}=1}} \omega_{rs}\mu_s^{k\to i}, \tag{S17}$$

and q_r^i is the one-node marginal posterior probability of node i belonging to group r defined previously in equation (S12). This probability can itself be calculated directly from the beliefs according to

$$q_r^i = \frac{\gamma_r}{Z_i} \exp\biggl(-\sum_k d_i d_k \sum_s \omega_{rs} q_s^k\biggr) \prod_{\substack{k\\ a_{ik}=1}} \omega_{rs}\mu_s^{k\to i}, \tag{S18}$$

with

$$Z_i = \gamma_r \sum_r \exp\biggl(-\sum_k d_i d_k \sum_s \omega_{rs} q_s^k\biggr) \prod_{\substack{k\\ a_{ik}=1}} \omega_{rs}\mu_s^{k\to i}. \tag{S19}$$

The belief propagation calculation involves choosing an initial set of values for the beliefs and the one-node probabilities (for instance at random in the interval $[0, 1]$), using them first to calculate new values of the q_r^i from Eqs. (S18) and (S19), and then using those values, plus the beliefs, to calculate new values of the beliefs from Eqs. (S16) and (S17). Then we repeat the procedure, iterating until the beliefs converge.

This gives a set of beliefs for the current values of the parameters γ, ω. Returning to the EM algorithm, we then use those values to compute improved estimates of the parameters from equations (S10) and (S11). To do this, we first need to calculate the two-node marginal probabilities q_{rs}^{ij} from the beliefs, which we do as follows.

Note that q_{rs}^{ij} appears only in the sum in the numerator of equation (S11) and that the sum involves only the values of q_{rs}^{ij} for node pairs i, j that are connected by an edge. (Those not

connected by an edge have $a_{ij} = 0$ and hence do not appear in the sum.) For pairs connected by an edge, q_{rs}^{ij} is by definition equal to

$$q_{rs}^{ij} = P(c_i = r, c_j = s|a_{ij} = 1, A') = P(a_{ij} = 1|c_i = r, c_j = s, A')\frac{P(c_i = r, c_j = s|A')}{P(a_{ij} = 1|A')}, \quad \text{(S20)}$$

where the parameters γ, ω are assumed given in each probability and A' denotes the set of elements of the adjacency matrix excluding a_{ij} (which is specified separately). But each term in this expression is now straightforward to write in terms of quantities we already know. The probability $P(a_{ij} = 1|c_i = r, c_j = s, A')$ is just the likelihood of the edge from i to j, which for our stochastic block model is

$$P(a_{ij} = 1|c_i = r, c_j = s, A') = d_i d_j \omega_{r,s} \mathrm{e}^{-d_i d_j \omega_{rs}}. \quad \text{(S21)}$$

Since ω_{rs} is typically very small, it is usually acceptable to neglect the exponential. (Recall that we are only interested in assigning each vertex to the highest-probability community, so small errors in the probabilities typically make no difference to the final answer.) And the probability that $c_i = r$ given A' is precisely the belief $\mu_r^{i\to j}$, so

$$P(c_i = r, c_j = s|A') = \mu_r^{i\to j}\mu_s^{j\to i}. \quad \text{(S22)}$$

The probability $P(a_{ij} = 1|A')$ is fixed by the requirement of normalization, meaning it can be calculated by stipulating that $\sum_{rs} q_{rs}^{ij} = 1$. The end result is

$$q_{rs}^{ij} = \frac{d_i d_j \omega_{rs} \mu_r^{i\to j}\mu_s^{j\to i}}{\sum_{rs} d_i d_j \omega_{rs} \mu_r^{i\to j}\mu_s^{j\to i}}. \quad \text{(S23)}$$

Substituting this value into equation (S11) now gives us our new value for ω_{rs}.

Our final, combined EM/belief propagation algorithm now consists of the following steps:

1. We choose initial values of the parameters γ_r and ω_{rs} for all r, s, for instance at random.
2. We choose initial values of the beliefs $\mu_r^{i\to j}$ and one-node marginal probabilities q_r^i, for instance at random.

3. We iterate the belief propagation equations (S16)–(S19) to convergence to give values for the beliefs $\mu_r^{i\to j}$ and the one-node marginal probabilities q_r^i.

4. We use these values to calculate the two-node probabilities q_{rs}^{ij} from equation (S23).

5. We use the one- and two-node probabilities to calculate improved estimates of γ_r and ω_{rs} for all r, s from equations (S10) and (S11).

6. We repeat steps 3 to 5 until the parameters and probabilities converge.

7. We assign each node to the community r for which its probability of membership q_r^i is highest.

2.4 Number of submarkets

When applied to the networks of heterosexual dating studied here, the algorithm of the previous section finds clear community structure. In fact, there are two different types of structure found, one essentially trivial, the other not. The trivial structure is a division between men and women. Almost all messages on the web site between heterosexual users looking for romantic relationships are between a man and a woman—well over 99%. Very few are between two men or two women. Our algorithm readily perceives this structure, reliably dividing the network into men and women without the need for us to identify the sexes explicitly. This "disassortative" structure is characterized by a matrix ω_{rs} of probabilities that has almost all of its weight off the diagonal (most connections are between different groups) and virtually none on the diagonal (connections between members of the same group).

In addition to this trivial structure, however, there is also the nontrivial group structure that we refer to as submarkets—the tendency of the population to break up into distinct communities of dating with relatively little message traffic between communities.

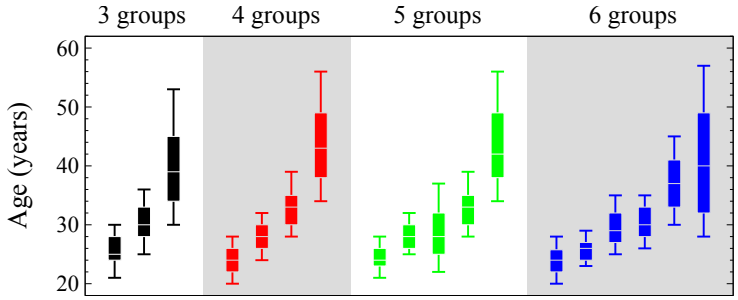

Figure S2: Box plots of the age ranges within submarkets for divisions of the New York City user population in to three, four, five, and six submarkets. For simplicity, men and women are combined in each submarket in this plot, but a similar pattern is seen when one examines the ages of men and women separately.

A practical upshot of this is that if we wish to divide our network into, say, four submarkets, we must actually instruct our algorithm to look for twice this number of communities (i.e., eight). If we do this, then it reliably finds four submarkets, each further divided into men and women.

In the calculations presented in the paper we chose to divide each city into four submarkets, but divisions into other numbers of submarkets would also be reasonable. To explore the effect of varying the number of submarkets we have performed divisions of the networks into various numbers of communities. Figure S2 shows the results of several possible divisions of the New York City network. (Similar patterns are seen in the other three cities.) The panels of the figure show the age distribution (men and women combined) for divisions into three, four, five, and six submarkets (which means six, eight, ten, and twelve communities in total, once the trivial division between men and women is factored in). As we can see, the primary effect of increasing the number of submarkets is to divide the population into more closely spaced age ranges, so that divisions into larger numbers of groups give a finer, more granular, picture of the market structure but the same overall behavior. As with all statistical analyses in which data are divided into bins, there is a balance to be struck between larger numbers of bins, which gives finer detail in the analysis, and smaller numbers of bins, which gives better statistics. Our choice of four

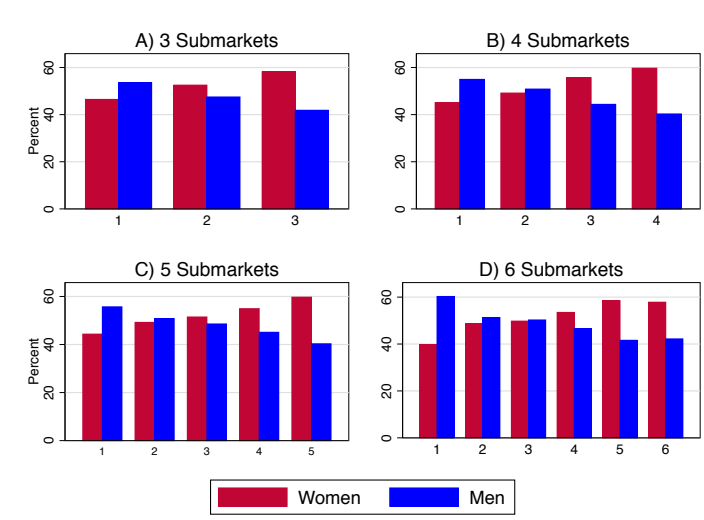

Figure S3: Fractions of men and women in each submarket for divisions of the New York City dating market into three, four, five, and six submarkets. The systematic pattern, seen in the Fig. 1B of the paper, in which the ratio of men to women becomes progressively more female-heavy as we move into the older submarkets, is duplicated in each case here, demonstrating that this is a general behavior, and is not particular to any one choice of number of submarkets.

submarkets per city gives a good picture of the overall behavior while maintaining sufficient statistical power for accurate analysis of the population within submarkets.

The systematic variation of the ratio of numbers of men and women among submarkets seen in Fig. 2B of the paper also extends to divisions into other numbers of submarkets, as shown in Fig. S3. As the figure shows, the pattern for the four-way division of Fig. 1B, whereby the sex ratio becomes progressively more female-heavy as we move into the older submarkets, is duplicated for divisions into three, five, and six submarkets as well.

3 Additional analyses and results

In the paper, we observe that minority women tend to be younger than white women in the same submarket, a trend that is particularly noticeable for black women. While the pattern holds across all of our four cities, it is most pronounced in Chicago. Here we provide additional details on the racial composition of Chicago users and insight into processes that give rise to the age differences we observe between white and black women in Chicago. We also examine whether the patterns observed in Chicago hold in New York, the other city with a sizable black population.

Figure S4 shows the mix of ethnicities for men and women in each Chicago submarket. The predominant group in all submarkets is whites, which reflects the overall composition of the Chicago user base. There is, however, systematic variation in the relative size of the minority population across submarkets. Black men and women are more prevalent in the oldest submarkets, which is surprising given that they are slightly younger, on average, than their white counterparts. One factor driving this is that the black women messaged by both black and white men are, on average, significantly younger than the white women messaged by men in the same submarket, and this phenomenon is most pronounced in the oldest submarkets. This tends to pull younger women into the older submarkets, and with them the men that they exchanges messages with. This helps explain not only why there is a surplus of black women in the oldest submarket, but also why these women are significantly younger, on average, than white women in the same submarket.

Figure S5 extends our analysis of age differences in messaging by submarket and race (Fig. 3 in the paper) to Boston and Seattle. The pattern is similar overall to that for New York and Chicago: age differences tend to be larger for first messages than for replies, and also larger in older submarkets. In submarket 4, for example, white men initiate contact with Asian women

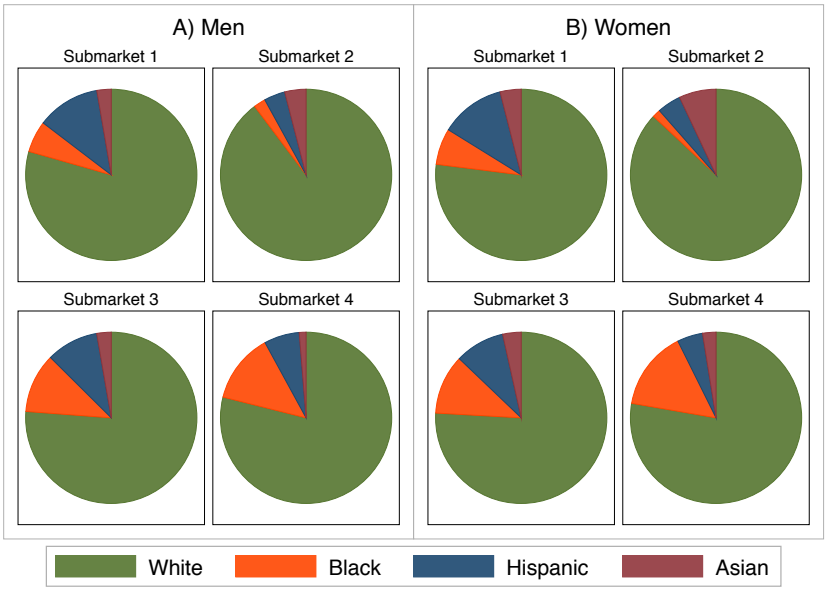

Figure S4: Racial composition of submarkets in Chicago. All submarkets are predominantly white, which is consistent with the overall composition of the Chicago market. However, despite the fact that whites are older, on average, than other race groups, they are disproportionately concentrated in submarket 2. Black users, especially black women, are over-represented in the oldest submarkets. Figure 3 in the text suggests one mechanism driving these patterns.

who are around 6 years younger than themselves on average, but receive replies from women who are only around 3.5 years younger. Also in line with the patterns for New York and Chicago, we see that within a given submarket non-white women tend to receive messages from older men than do white women; this is especially true in submarket 4.

There are, however, also some striking differences between the results for Seattle and Boston and those for New York and Chicago. In Boston and Seattle, women in submarket 4 (and for Seattle submarket 3 as well) display little tolerance for overtures from much older men. Note how in these cities women's replies are predominantly to men of similar age to themselves, despite the fact that men are messaging significantly younger women. Black women in Seattle

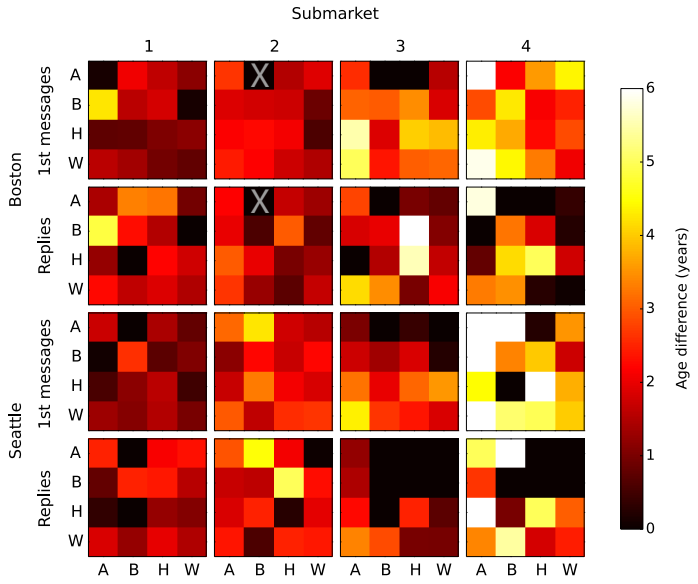

Figure S5: Mean difference in years between the age of men of varying races and the women they message in Seattle and Boston, by race of women and submarket. Race is coded as: A = Asian, B = Black, H = Hispanic, and W = white. The first two rows show the average age difference for, respectively, all initial messages sent in Boston and those that received a reply; and the bottom two rows show the same patterns for Seattle. We observe zero instances in Boston where black women receive messages from Asian men in submarket 2, so these cells are marked with an *X*.

for example are receiving overtures from black men about 3.5 years older than themselves on average, but reply primarily to men of about their own age. Notable exceptions to this behavior are messages from Asian men to Asian women, and from Hispanic men to Hispanic women, which appear to receive replies despite large average age differences.

References

1. Bruch, E. E. & Newman, M. E. J. Aspirational pursuit of mates in online dating markets. *Science Advances* **4**, eaap9815 (2018).

2. Hitsch, G. J., Hortaçsu, A. & Ariely, D. Matching and sorting in online dating. *American Economic Review* **100**, 130–163 (2010).

3. Lin, K.-H. & Lundquist, J. Mate selection in cyberspace: The intersection of race, gender, and education. *American Journal of Sociology* **119**, 183–215 (2013).

4. Lewis, K. The limits of racial prejudice. *Proc. Natl. Acad. Sci. USA* **110**, 18814–18819 (2013).

5. Nowicki, K. & Snijders, T. A. B. Estimation and prediction for stochastic blockstructures. *J. Amer. Stat. Assoc.* **96**, 1077–1087 (2001).

6. Bickel, P. J. & Chen, A. A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106**, 21068–21073 (2009).

7. Karrer, B. & Newman, M. E. J. Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83**, 016107 (2011).

8. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications. *Phys. Rev. E* **84**, 066106 (2011).

9. Decelle, A., Krzakala, F., Moore, C. & Zdeborová, L. Inference and phase transitions in the detection of modules in sparse networks. *Phys. Rev. Lett.* **107**, 065701 (2011).

10. Yan, X. *et al.* Model selection for degree-corrected block models. *J. Stat. Mech.* **2014**, P05007 (2014).