sociological science

Rejoinder to Lewis

Ashton Anderson,[a] Sharad Goel,[a] Gregory Huber,[b] Neil Malhotra,[a] Duncan J. Watts[c]

a) Stanford University; b) Yale University; c) Microsoft Research

In a response to our article "Political Ideology and Racial Preferences in Online Dating," Kevin Lewis (2015) has raised concerns about various aspects of our analysis of user behavior on a national online dating site. The core of Lewis' critique is that we do not report specifics of the sample, the website, and the analysis. Lewis also details numerous smaller points with respect to details of the analysis. The goal of this response is not to reply to every point that Lewis makes. As with any research project, our design decisions require tradeoffs, and readers should evaluate the evidence on their own. Rather, we use this space to make broader points about the general value for sociology of research using online data and how such research should be practiced.

Analytical Reporting and the Value of Mixed Methods

A recurrent theme in Lewis's critique is that we should have reported additional details about our data and how we analyze it. This desire is understandable, but fails to acknowledge an important tradeoff that social scientists face between detail and insight. As with any form of data-driven research, reporting every minute detail of how the data were generated and analyzed would prevent readers from seeing the forest for the trees—a point that holds for interview research or research based on administrative data as much as for online data. Insofar as all data analysis is reductionist, the key question is not whether we include every detail, but instead whether we omit some key fact about how our data were generated or how our data analysis is conducted that would undercut the inferences we present. No doubt we could have burdened the reader with many more such details, but we believe that we included those necessary to understand and evaluate our claims.

Lewis also states that "it is important that we acknowledge both the advantages and limitations of this research." We agree. Again, as with all types of data and research designs, ours has strengths and limitations. The point of our article is not that researchers should *only* analyze online data, but that these types of data bring unique strengths and limitations to important research questions. Lewis seems to privilege traditional approaches of measuring homophily, such as post-match surveys of formed relationships. However, as we argue, those data have different and substantial limitations: 1) stated responses in surveys may be subject to social desirability bias; 2) even in the absence of social desirability bias, people may not be completely aware of their own biases, and therefore stated and revealed preferences might diverge; 3) by measuring attributes of individuals *after* they have matched,

Citation: Anderson, Ashton, Sharad Goel, Gregory Huber, Neil Malhotra, and Duncan J. Watts. 2015. "Rejoinder to Lewis" Sociological Science 2: 32-35.

Received: November 13, 2014

Accepted: November 17, 2014

Published: January 21, 2015

Editor(s): Jesper B. Sørensen

DOI: 10.15195/v2.a3

preference-based homophily is conflated by post-match convergence; and 4) because the potential pool of dating partners is unknown, homophily could arise solely from structural features of a person's social environment. We readily acknowledge weaknesses in our own dataset but contend that evidence for preference-based homophily is strengthened when online data supplement existing data from surveys. Sociology has always relied on evidence from multiple methods; our article was intended in the same spirit.

Sampling

We now address some minor points specific to our study. First, Lewis criticizes us for not providing details on the set of users we started with before winnowing the sample down to one where we have full data. Such reporting, however, would be of minimal utility as we make no claims that our data are representative of any population—nor would that be any more true of the full sample of users than the winnowed sample. Accordingly, comparing the analyzed sample to the full sample of users would do little to speak to the overall representativeness of the data. Lewis seems to suggest that survey data do not have these problems; yet given that response rates to surveys are not 100 percent, survey data do in fact have similar issues in that not all individuals agree to be surveyed and we do not observe full data for everyone who was asked to participate in a survey. Lewis also hypothesizes that people who are more open to reporting demographic information (e.g., income) would be more likely to admit they have same-race preferences. However, if this conjecture is accurate, our sample population would exhibit a *smaller* difference between stated and revealed preferences, meaning that the differences that we do find represent lower bounds. Finally, Lewis criticizes our decision to analyze only blacks and whites. However, this criticism is not about research design or methodology so much as a disagreement with our specific analytical choices. Our decision stems from the methodological challenges of studying Hispanics and Asians in this context, including relatively small sample sizes and ambiguous self-identification of race. In theory, however, such data could be used to study these subpopulations, and we encourage other scholars to do so. The key point is that our focus in no way invalidates the analysis done *among* blacks and whites.

Dating Site Algorithms

One of the major hurdles of working with online dating sites is that the potential partners presented to users are not random. This is of course the value of "match-making": the consideration set is filtered to reduce search costs. Lewis criticizes our study by claiming that the algorithm could have substantially affected the results. Algorithmic confounding is a potential problem for data generated by web platforms that have an interest in optimizing user experience. In our case, however, it is much less of a problem than Lewis suggests. Unlike many previous analyses, we have a very good understanding of how the dating site we are studying operates, including specific knowledge of the partner recommendation algorithm, which is extremely simple and transparent compared to more sophisticated matchmaking services available today. Note that this feature of our work stands in contrast to

research that has analyzed data from sites such as OKCupid, where the algorithm is quite opaque due to the need to protect intellectual property.

As we state in the article, the algorithm prioritizes profiles for which people have a "must-have" preference. To account for this, we implement a model-based approach (also described in the original article) that attempts to reconstruct the set of potential partner choices available to the user and then analyzes how people select from that set. We acknowledge that this approach involves some assumptions, but this is no different than previous studies that have analyzed online dating data.

The main breakthrough of our article, however, is that we can directly control for the impact of the algorithm: When we restrict our attention to respondents who selected "nice-to-have" versus "no preference," the algorithm does not play any role in determining the pool of potential partner profiles. And our analysis of this subsample yields similar substantive conclusions as our model-based approach—we find that people who state "no preference" exhibit same-race revealed preference, but that these levels of same-race revealed preference are lower than for those who selected "nice-to-have." Lewis does not have a substantive critique of this analysis; he only states that this sample is different than the one where people selected "must-have." That is true, and is a potential limitation, but Lewis does not properly acknowledge as a major advantage of our data the fact that we are able to remove the partner suggestion algorithm as a confounding explanation for our results. Further, though we make the data analysis more manageable by limiting potential partners based on geographic and demographic constraints, these restrictions have nothing to do with the algorithm, as Lewis implies.

Summary

In summary, Lewis points out that: 1) our data could have been better; 2) we could have made different decisions in our research design; and 3) as a consequence, our results may have limited generalizability. Although we have specific disagreements with some of Lewis's criticisms (most importantly that we did not account for possible bias introduced by the matching algorithm), we don't disagree with these general remarks. That said, we also feel that similar remarks could be applied to essentially all empirical papers, including those that rely on traditional sources of data. If we had advocated web data as a panacea to empirical social science or had claimed that our specific data and analysis approach were free of all possible bias, then Lewis's critique would be a useful warning. But we did not make any such claims. Nor do his specific disagreements undermine the substantive conclusions of our article.

We have shown evidence of preferences to date people of the same race, and that these preferences correlate with political ideology. Furthermore, while there is often a divergence between stated and revealed preferences, these gaps are no different for liberals and conservatives. Our data have limitations, but so do more traditional survey data. The point is that our data have unique strengths and limitations, which increases our confidence that people have both explicit and implicit preferences for same-race partners as shown in the survey data. Further, even among studies of online behavior, our study stands out because we are able both to pinpoint the

algorithm that produces the set of partner choices available to users and to measure directly the effects the algorithm has on our results.

As our design has improved on previous efforts, we look forward to future analyses of online behavior that improve on the weaknesses of our research design, including studies that use more traditional datasets. The advancement of knowledge in sociology has depended on diverse methodological approaches and sources of data. Analyzing the digital footprints people leave as part of their online behavior will continue to be an important approach in the study of homophily and other important substantive questions.

References

Lewis, Kevin. 2015. "Studying Online Behavior: Comment on Anderson et al. 2014" Sociological Science 2: 20-31. http://dx.doi.org/10.15195/v2.a2.

Ashton Anderson: Department of Computer Science, Stanford University. E-mail: ashton@cs.stanford.edu.

Sharad Goel: Department of Management Science and Engineering, Stanford University. E-mail: scgoel@stanford.edu.

Gregory Huber: Department of Political Science, Yale University. E-mail: gregory.huber@yale.edu.

Neil Malhotra: Graduate School of Business, Stanford University. E-mail: neilm@stanford.edu.

Duncan J. Watts : Microsoft Research. E-mail: duncan@microsoft.com.