sociological science

Supplement to:

Arvidsson, Martin, Peter Hedström, Marc Keuschnigg. 2025. “Wide Social Influence and the Emergence of the Unexpected: An Empirical Test Using Spotify Data” Sociological Science 12: 715-742.

Wide Social Influence and the Emergence of the Unexpected: An Empirical Test Using Spotify Data

Appendix

Martin Arvidsson[1,*], Peter Hedström[1,*], and Marc Keuschnigg[1,2,*]

[1]The Institute for Analytical Sociology, Linköping University, Sweden.
[2]Institute of Sociology, Leipzig University, Germany.
[*]Corresponding authors:
martin.arvidsson@liu.se, peter.hedstrom@liu.se, marc.keuschnigg@liu.se

S1 Data and Descriptives

Table S1 lists the number of treatments and adoptions among treated users in the four user subsets for which we report separate treatment effects. We define subpopulations by the degree of user–artist compatibility (low, high) and by the taste overlap (weak, strong) between Ego and the Alter from whom exposure originated. Our empirical analysis rests on data for Spotify users who use the platform's social networking function. This user group may differ in some respects from those who do not make use of this function.

TABLE S1
Number of treatments and adoptions among the treated users in the four user subsets.

Artist compatibility	Ego–Alter taste overlap	*N* Adoptions	*N* Treated
Low	Weak	950	3,848,649
Low	Strong	975	2,998,057
High	Weak	4,043	3,814,266
High	Strong	12,240	4,931,458

Our empirical analysis focuses on artists whose popularity rank on Spotify spans 250–5,000. The lower bound avoids our results being dominated by the extreme tail, who account for a disproportionate number of treatments and adoptions in the raw data.

Excluding the superstars is important not only to ensure our estimates better reflect the broader population of artists on Spotify, but also because our primary interest lies not in the success of the already established, but in the spread of novel artists. The upper bound is set for practical reasons: as adoptions in specific time windows are rare, it ensures a sufficient number of treatments. To ensure both sufficient coverage and a focus on novelty, we apply an additional filter: among artists ranked 250–5000, we include only those to whom at least 0.5% (and at most 10%) of users were exposed during the observation period (January 1–November 23, 2016). Processing all artists that meet these criteria, we are after the statistical matching left with 1521 artists of satisfactory balance (see section S3 below), totaling 15,592,430 treatments. Including the most established artists does not change the qualitative nature of our results—the moderation effects follow the same pattern, but the overall magnitude of effects is lower (see Figure S5B below). Including artists with lower popularity is difficult due to rare outcomes and the increasing challenge of achieving balance with fewer treatments.

TABLE S2
Sociodemographic variables.

Variable	Description	Mean	SD	Median	Min	Max
Number of Alters	Users Ego follows (outdegree)	2.54	3.53	2	1	415
Activity level	Number of songs added	128.87	454.06	20	0	64,810
Experience	Days since registration	1,303.64	416.37	1402	59	2,920
Gender	Female	0.42				
	Male	0.41				
	No gender information	0.17				

Our data include a limited set of sociodemographic variables (Table S2). Ego's *number of Alters* (outdegree) averages 2.5; 15% of Egos follow at least one Alter also followed by another sampled Ego. *Activity level* is the average number of songs a user added during the observation period. As a proxy for user *experience*, we use days since registration, measured at the end of the observation period. We infer each user's *gender* from registered first names using the R package `gender` (U.S. Census and Social Security data); names that cannot be classified as female or male (17%) are mean-imputed using the sample mean of the binary indicator.

S2 Topic Model

Estimation and Inference

To estimate the LDA model, we use WarpLDA (Chen et al. 2015) as implemented in the R package text2vec (Selivanov et al. 2020).[1] For computational feasibility, we did not estimate the LDA on the full sample of 30 million playlists. Instead, we estimated the model on a random sample of 6 million playlists—large enough to capture the richness of the taste patterns in the full sample. To allow for evolving music taste, we infer users' topic proportions on a monthly base, considering all songs added each month, and re-aggregate to arrive at temporal music taste profiles. We constructed users' monthly taste profiles based on topic proportions at the level of a user's collapsed playlists, ensuring computational feasibility.

Selecting the Number of Topics K

For LDA, as with most dimensionality reduction and clustering techniques, the number of dimensions (topics) is a hyperparameter that needs to be specified prior to estimation. Following standard practice (Blei et al. 2003), we let the hold-out likelihood guide the choice of the number of topics: we (1) partitioned the data into a training set (90%) and a test set (10%), (2) fitted the LDA model on the training set using different numbers of topics, and (3) scored how "surprised" each model was by the held-out data—as measured by "perplexity," a normalized version of likelihood—given the estimated parameters of each model. The lower the perplexity, the better the generalized performance of the model. Figure S1 shows that the hold-out fit improves drastically as the number of topics increases from 10 to 100. Although the fit keeps improving with the number of topics, the improvements are marginal beyond 300 topics, and we selected K=300 as the number of topics used for the main analysis.

In choosing K, we considered both the relative improvement in fit for each increment in K as well as the fact that a larger K leads to more specialized topics and thus a more granular measurement of music tastes. We interpret topics as subgenres or temporal epochs such as "East coast hip hop," "death metal" or "80s pop." Granularity improves matching to better control potential confounding influences, including automated recommendations. For the purpose of matching, there is generally no danger in using too many topics, except that achieving balance on all dimensions becomes gradually more difficult as they grow in number (Roberts et al. 2020). For smaller K, topics merge and the confounding dimensions of music taste may be washed out by being combined with other non-confounding dimensions.

[1] We have also considered the parallelized Gibbs sampling implementation in MALLET (McCallum 2002), which provided consistent results.

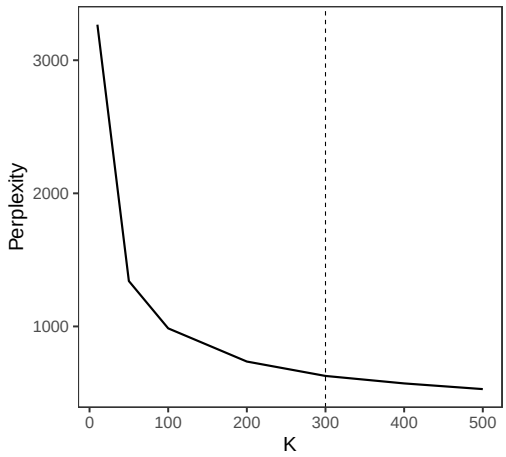

FIG. S1. Hold-out fit on the 10% test set, as measured by perplexity, for LDA models with different numbers of topics.

For larger K, topics become too granular, making it difficult to find good matches among increasingly diversified users. For the purpose of causal inference, a middle course is thus important (Roberts et al. 2020). We have replicated our analysis using both fewer (200) and considerably larger (1000) number of topics, and the results are highly consistent.

Artist Compatibility and User Homophily

We compute artist compatibility, the degree to which an artist fits into a user's taste profile, as a function of the user's topic proportions and the artists' topic weights. Letting U_i represent user i's K-dimensional topic profile and A_j artist j's weights to each topic, we calculate the compatibility $C_{i,j}$ between user i and artist j as

$$C_{i,j} = \sum_{k}^{K} U_{ik} A_{jk}.$$

Both U_{ik} and A_{jk} are restricted to values between 0 and 1, and $\sum_k^K U_{ik} = 1$ and $\sum_k^K A_{jk} \leq 1$, so that $C_{i,j}$, the sum of their products, also ranges from 0 to 1, with higher values representing greater compatibility between user and artist. To account for the fact that the sum of artists' topic weights, $\sum_k^K A_{jk}$, varies between artists (due to differences in absolute adoption frequencies, i.e., artist popularity), we consider the artist-normalized version

$$C^*_{i,j} = \sum_{k}^{K} \frac{C_{i,j} - \bar{C}_j}{\sigma_{c_j}},$$

where $\bar{C}_j$ is the average compatibility score for artist j and σ_{c_j} is the standard deviation of j's compatibility scores. We categorize artist into "low" (percentiles 1–50) and "high" compatibility (percentiles 51–100).

To measure the taste overlap between Egos and Alters, we use the cosine similarity score. Letting X and Y represent two vectors, or two taste profiles, we compute the cosine similarity H between X and Y as

$$H_{X,Y} = \cos(X,Y) = \frac{\sum_i^n X_i Y_i}{\sqrt{\sum_i^n X_i^2 \sum_i^n Y_i^2}}.$$

The resulting similarity scores can range between –1 (exactly opposite) and 1 (identical). Topic proportions are restricted to be non-negative (≥ 0), so the resulting homophily scores lie in the interval $[0,1]$. Because the taste overlap between Ego and the treating Alter is, by definition, only defined for treated users, we instead use the average taste overlap between Ego and all their Alters—which, in contrast, is well-defined for both treated and control users. This approximation makes our results conservative—with respect to effect differences between taste-overlap categories—for the reason that the dyadic overlap between Ego and the treating Alter is smoothed out by considering the average of all Alters. We group homophily scores into two categories, differentiating "weak" (percentiles 1–50) and "strong" (51–100) taste overlap between Egos and Alters. We chose cut-points that aim to distribute statistical power equally across the categories.

S3 Propensity Score Model

The statistical matching proceeds as follows: First, we estimate a logistic regression model separately for each artist who serves as a treatment using the treatment indicator as the binary dependent variable (1 for each user being exposed to the focal artist, 0 otherwise). We condition the models on each level of user-artist compatibility (low, high) and taste overlap (weak, strong) between Ego and the Alter. By estimating separate artist-specific logistic regression models for each of the 2 (compatibility) × 2 (homophily) = 4 combinations of the moderators of social-influence effects, we control for potential differences in balancing across user subpopulations defined by different degrees of compatibility and homophily (Green and Stuart 2014). The matching variables used to find statistical twins are users' music taste, their activity level on the platform, and a set of sociodemographic variables (see Table S2).

The number of topics relevant for predicting exposure to treatment is likely to be small compared to the total number of topics—a property often referred to as "approximate sparsity" in the literature on high-dimensional causal inference (e.g., Belloni 2014). We use

a forward-selection procedure to determine which variables (topics) should be included in each propensity score model (for details, see "Balancing" below). It is worth emphasizing that although this implies that we use only a subset of variables in the estimation of each propensity score model, all variables are ultimately considered in the evaluation of balance. This way, we utilize the granularity of the high-dimensional topic space while ensuring efficient estimation of the propensity score models.

Second, we pair each treated user with k counterfactual users who (1) never previously had adopted the artist in question, (2) up until the end of the adoption window of 1 week never were treated with that artist, (3) belonged to the same artist compatibility and Ego-Alter homophily categories as the treated user, and (4) minimized the absolute difference in the temporal propensity score of receiving the treatment. Actual k varies across Egos depending on the number of high-quality counterfactuals available, and we determine the quality threshold based on a propensity caliper of 0.1 times the standard deviation of the treatment group's monthly propensity scores. We sample counterfactual users with replacement, such that users can re-appear as controls for multiple treatments (Easley et al. 2020).

Balancing

We apply the following procedure separately for each artist who served as a treatment and for each user subset (within each artist) for which we report separate treatment effects, i.e. the 4 user subpopulations defined by user–artist compatibility (low, high) and Ego–Alter taste overlap (weak, strong). This leads to 1521 artists × 2 artist-compatibility levels × 2 user-homophily levels = 6,084 matched samples.

1. Calculate the absolute standardized difference in means—prior to matching—for all variables and select the variables with a standardized difference in means greater than 0.1.

2. Estimate the propensity score matching model using the selected variables.

3. Match each treated user with k untreated users (within a 0.1 × standard deviation caliper), and re-compute the absolute standardized difference in means in the resulting matched sample. For variables that were included in the propensity score model of the previous step, but remain above 0.1, add a one-degree-higher polynomial term. For variables that become imbalanced but were not included in the previous step, include a first-degree (i.e., linear) coefficient.

4. Repeat steps 2–3 until all dimensions have a standardized difference in means below 0.1, or for a maximum of 5 iterations. If, at iteration 5, any variable has a standardized

difference greater than 0.11, the matched sample in question is excluded from further analyses due to insufficient balancing between treated and untreated users.

We considered alternatives to the standardized difference-based rule that are computationally more intense, such as lasso regression, and we found them achieving similar levels of balance.

Temporal Scale

Another detail of the matching procedure concerns the temporal scale at which we estimated the models and predicted propensity scores. For estimation, we included all users who at some point were available for treatment (the risk set contains the users who had not previously adopted or been exposed to the treatment artist in question) in the observation period. We considered one observation per user, representing the last recorded month the user had been available for treatment. The topic dimensions then reflect the user's temporal taste profile and activity level as of that month. For any specific artist, the great majority of users were available for treatment in the last observation month (i.e., they never were exposed to or never adopted the treatment artist in question). Following estimation, we predict monthly propensity scores (aligned with the monthly-inferred taste profiles), allowing for individual-level changes in the propensity of treatment throughout the observation period.

Weighting

Given time-windowed treatments and outcomes, artist-level treatment status, and month-conditioned tastes, we define weights at the matched-group level and assign equal weights to controls within each matched group under variable-ratio matching with replacement. For each treated Ego–artist–time matched group with k counterfactuals, each counterfactual receives weight $1/k$ (weights sum to 1 within the group). If a counterfactual user appears in multiple groups, they contribute one $1/k$-weighted observation per appearance. Uncertainty is quantified with a two-stage bootstrap that accounts for the nested structure of the data (matched groups within artists; see Section S6).

S4 Balance Diagnostics

To examine the balance between the treatment and control group in the matched samples, we compute the standardized mean difference d for all confounders. We perform this balancing test separately for each artist serving as a treatment and for all user subsets within

each artist for which we estimate separate treatment effects. As a point of reference, we also calculate balance scores for the full sample, showing how imbalanced the data are when confounders are not adjusted for.

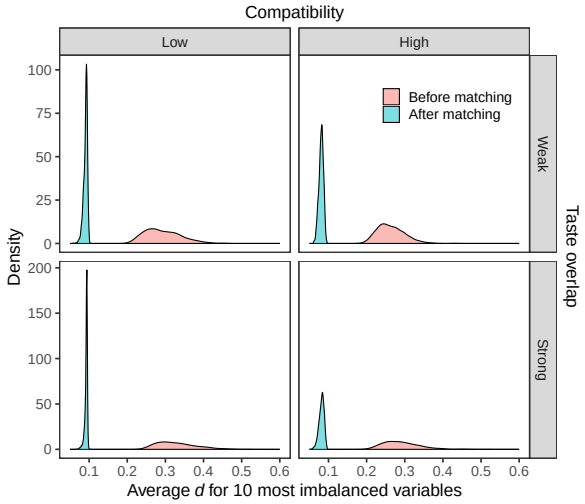

FIG. S2. Average standardized difference in means for the 10 most imbalanced variables for each of the 6,084 artist × user-subset combinations, plotted as densities over artists.

To effectively show the balance for 1521 artists × 4 user subsets × 604 variables (300 Ego topic proportions and 300 average Alter topic proportions, plus 4 sociodemographic variables about Ego; see Table 2), we create plots summarizing results in terms of distributions over artist-specific statistics that are conditional on the user subsets. Figure S2 displays the average d among the top 10 most imbalanced variables for each artist × user-subset combination (capturing the imbalance of the top 2% most imbalanced variables for each combination). To derive each result, we (1) compute d for all 604 variables, (2) rank-order the variables according to d, and (3) calculate an average d of the 10 variables with the largest d. We observe that, without adjusting for confounding, the average d among the top 10 most imbalanced dimensions clearly exceeds the conventional thresholds of 0.10 (Austin 2009) and 0.25 (Stuart and Rubin 2008; Stuart 2010) for the great majority of artist × user-subset combinations. By contrast, in the matched samples generated by our taste-based matching procedure, the average d for the top 10 most imbalanced dimensions drops to around or below 0.10 for all combinations.

Complementing this evaluation, Figure S3 plots the number of dimensions that exceed the conventional thresholds of d for each artist × user-subset combination. We observe that,

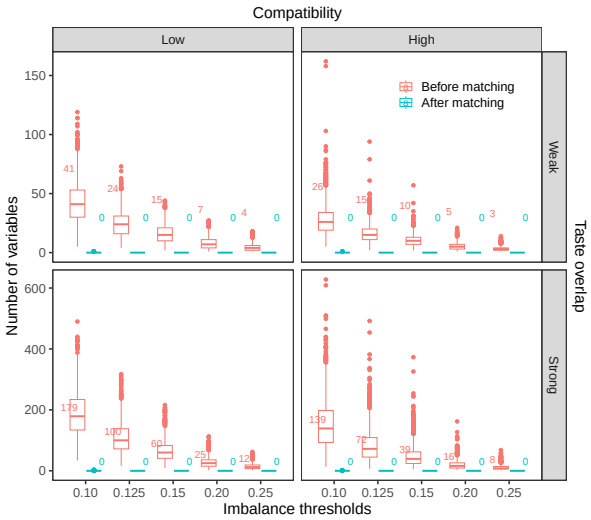

FIG. S3. Number of variables that exceed conventional standardized difference thresholds, plotted as boxplots over artists for each of the 6,084 artist–user subset combinations. The numbers show the median number of variables exceeding the thresholds in each user subset.

without adjusting for confounding, dozens of variables are considerably imbalanced (with d exceeding both 0.1 and 0.25) for most combinations. After taste-based matching, however, the typical case instead is that either no (the median is 0) or very few dimensions exceed the conservative threshold of 0.1 and that all dimensions remain well below 0.25. This balancing is achieved while retaining the great majority of treatments. In sum, we find that our taste-based matching procedure can create well-balanced matched samples with respect to music taste, listening activity, and various sociodemographic user properties.

S5 Identification

In the context of Spotify, the central cause of both cultural choice and tie formation is musical taste. Thus, to identify the effect of social influence on the probability of adopting new music, users' musical taste must be adjusted for. While musical taste is unobserved, we observe traces of users' listening behavior, and these can be viewed as proxies generated from the unobserved confounder. Figure S4 shows a directed acyclic graph (DAG) that describes our assumptions about the causal process (Morgan and Winship 2014). Let Z_i represent the musical taste of user i, X_i their adoption traces, $Y_{i,t}$ the adoption status of the artist in question at time t, and $A_{i,j}$ whether user i follows user j.

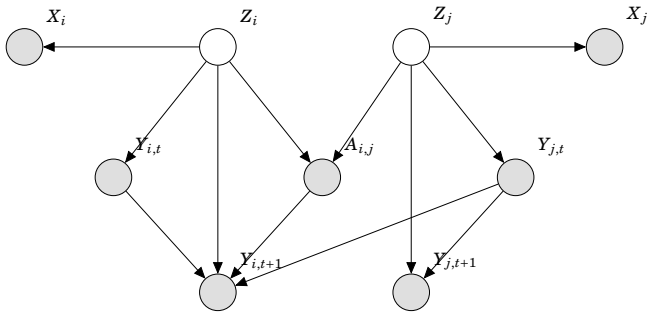

FIG. S4. Directed acyclic graph representation of assumed causal relationships.

Using this representation, the treatment effect δ that we seek to identify is the effect that $Y_{j,t}$ exerts on $Y_{i,t+1}$:

$$\delta = P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 1\right)\right) - P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 0\right)\right)$$

This effect can be identified—and the do-operators removed—if there are no open back-door paths between $Y_{j,t}$ and $Y_{i,t+1}$. However, because Z_i and Z_j are unobserved, and the collider $A_{i,j}$ is observed, the three together open a back-door path between $Y_{j,t}$ and $Y_{i,t+1}$ (Pearl 2009). Hence, we are unable to remove the do-operators from the following equation, and the effect remains unidentifiable:

$$\delta = P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 1\right), Z_i, Z_j\right) - P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 0\right), Z_i, Z_j\right)$$

This conclusion, however, does not consider the variables X_i and X_j. As the literature on proxy variables demonstrates (Kuroki and Pearl 2014; Louizos et al. 2017; Peña 2020), such back-door paths can be closed if there are descendants of the unobserved confounders that provide good approximations of Z_i and Z_j. Accepting the causal assumptions depicted in Figure S4, the identifiability of the social-influence effect thus relies on the extent to which past listening patterns (X_i and X_j) are good proxies for musical taste (Z_i and Z_j). If, as we have argued, X_i and X_j are reasonably good proxies of Z_i and Z_j, the back-door path closes, and the effect of $Y_{j,t}$ on $Y_{i,t+1}$ can be identified:

$$\delta = P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 1\right), X_i, X_j\right) - P\left(Y_{i,t+1} = 1 \middle| do\left(Y_{j,t} = 0\right), X_i, X_j\right) =$$

$$P\left(Y_{i,t+1} = 1 \middle| Y_{j,t} = 1, X_i, X_j\right) - P\left(Y_{i,t+1} = 1 \middle| Y_{j,t} = 0, X_i, X_j\right)$$

S6 Regression Results

In this section, we provide (1) additional details on the analysis performed on the matched sample, including our computation of standard errors, (2) information about the regression models underlying the empirical results presented in the main text, and (3) additional results from supplementary matching specifications that provide baselines to compare with the reported results of the high-dimensional matching approach.

To estimate the treatment effect of social influence, we compare the adoption behavior among treated users with the adoption behavior among counterfactual users who never previously had adopted the artist in question and up until the end of the 7-day adoption window never were treated with that artist.

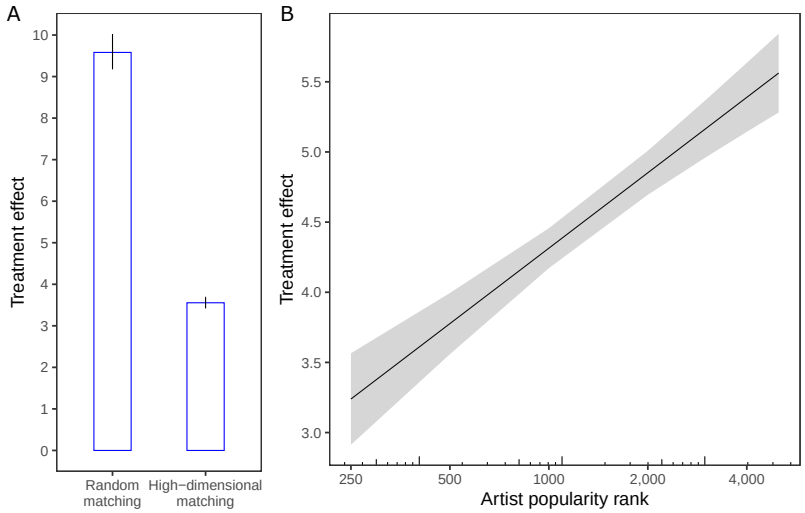

FIG. S5. (*A*) Estimates of the population-level treatment effect δ, based on random matching and matching on music taste. (*B*) Regression of artist-specific treatment effects against artists' download rank among sampled Spotify users (p<0.001; 95% confidence intervals).

Figure S5A compares the estimated treatment effect from high-dimensional matching on music taste (our primary analysis) with a random matching baseline. With high-dimensional matching, the estimated population-level treatment effect δ is 3.6 (±0.2, 95% bootstrap confidence interval), suggesting that exposure leads to a 3.6-fold increase in the average adoption likelihood. Random matching, where we randomly select untreated users as the control group—i.e., without adjusting for homophilic selection or correlated behaviors—results in a substantially higher estimate of the social-influence effect (9.6 ±0.4).

Not adjusting for homophilic selection thus inflates the estimated effect by approximately 167% compared to high-dimensional matching.

Figure S5*B* plots the result from regressing artist-specific treatment effects (δ_a, on the y-axis) against artist recognition (measured by download rank in our sample, on the x-axis). The negative slope (p<0.001) indicates that social influence is considerably stronger for less recognized artists, suggesting it is more efficacious for unfamiliar content. Interpreting artist recognition as inversely related to novelty, this finding corroborates our result that social influence is particularly effective in inducing adoptions outside individuals' usual repertoires.

Note that for the moderation analyses that speak to the core arguments of the article, absolute effect sizes are of secondary relevance (note also that the absolute effect sizes also depend on the length of the adoption-time window within which we follow up on the treatment event as well as on our intention-to-treat design). Much more important are the relative effect sizes, comparing social-influence effects at different levels of user–artist compatibility and Ego-Alter taste overlap.

Estimated Fixed-Effects Models

Table S3 summarizes the regression results underlying our empirical results (Figures 2 and 3). Model 1, considering the treatment indicator (and artist fixed effects), estimates the overall average treatment effect on the treated ($\delta = 3.6$) presented in Figure S5. Exponentiating the logit coefficients approximates the treatment effects that in the manuscript we report as relative risks derived from marginal predicted probabilities (for the population-level treatment effect of model 1, this means $\exp(1.3) \approx 3.6$). Model 2 expands on this analysis by incorporating user–artist compatibility and estimating treatment effects conditional on compatibility, presented in Figure 2. Model 3 further includes variables capturing the taste overlap between Egos and their Alters and estimates the treatment effect conditional on both compatibility and taste overlap, presented in Figure 3.

Non-Parametric Cluster Bootstrap

To quantify the uncertainty of the estimated treatment effects reported in the main text, we implemented a non-parametric cluster bootstrap procedure (Davison and Hinkley 1997) that accounts for both the artist-level variation and the within-artist variation in the matched groups (each consisting of one treated user and k untreated counterfactual users). The core of the implementation is similar to Davison and Hinkley's (1997) "two-stage bootstrap," except that the matched groups rather than individual observations (users) are resampled

TABLE S3
Estimates of the logistic fixed-effects regressions.

	Model 1		Model 2		Model 3	
	Estimate	SE	Estimate	SE	Estimate	SE
Untreated	ref.		ref.		ref.	
Treated	1.268	0.020	1.419	0.042	1.290	0.097
Artist compatibility: low	ref.		ref.		ref.	
Artist compatibility: high			2.005	0.030	1.711	0.046
Treated × high compatibility			–0.165	0.036	–0.271	0.089
Taste overlap: weak					ref.	
Taste overlap: strong					0.012	0.035
Treated × strong overlap					0.270	0.079
High compatibility × strong overlap					0.467	0.028
Treated × high compatibility × strong overlap					0.058	0.074

NOTE.—Binary dependent variable: adoption (0/1). All models include 1,521 artist fixed effects; bootstrap standard errors (SE) reported. In each model, the effective sample size N is 31,184,860.

within clusters (artists). This is our resampling procedure in detail:

1. Resample the 1521 unique artists at random with replacement.
2. For each of the 1521 sampled artists, resample the matched groups at random with replacement.
3. Based on the bootstrap sample created by steps 1–2, estimate a logistic fixed-effects regression, derive marginal predictions, and compute the desired risk ratios as well as their differences (to test for differences).
4. Repeat steps 1–3 10,000 times.
5. Compute confidence intervals of the different bootstrapped statistics by evaluating the quantiles of the bootstrapped sample. Significance testing is then performed by checking overlap, at the given significance level, with the value implied by the null hypothesis.

S7 Robustness Checks

Treatment Validity

Although most users only follow a small number of other users, we cannot be certain that all users were aware of their treatments. Hence, we interpret our treatments to be of the "intention-to-treat" kind (Dunning 2012), as is common in online studies of social influence (e.g., Bond et al. 2012; Aral and Nicolaides 2017; Eckles and Bakshy 2020).

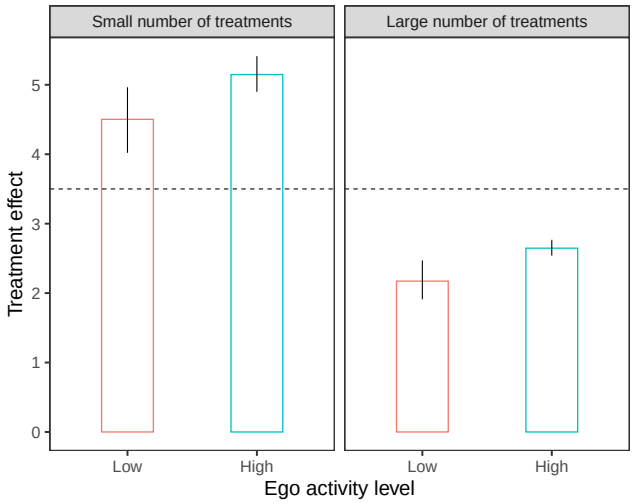

FIG. S6. Average treatment effect for users of different levels of platform activity (x-axis categories) and for users who received few, some, or many treatments in the three months prior to treatment, our measure of information overload (panels). The dashed line represents the treatment effect of 3.6±0.2 based on all user types and reported in the main analysis.

Actual exposure to the treatment can be expected to be more likely for high-activity users than for users who spend less time on the platform. It can also be expected to be less likely for users who in the three months leading up to the focal treatment had been exposed to a larger number of new artists through their peers. To further assess treatment validity, Figure S6 distinguishes between users with different likelihoods of compliance to treatment assignment. Indeed, the average estimated treatment effect is larger for high-activity users that lie in the 75th percentile (and above) when it comes to adding new songs to their profiles compared to less active users (<75 percentile). Estimated treatment effects are also smaller for users overloaded with exposures (≥75 percentile) than for users exposed to fewer treatments (<75 percentile) in the three months leading up to the focal treatment.

Note, however, that such subgroupings can introduce biases because users self-select into certain activity and networking levels, making it difficult to attribute subgroup differences in effect sizes solely to differences in the awareness of treatments. In our main analysis, we thus prioritized unbiased estimates over larger effect sizes and presented the more conservative estimates that rest on the intention-to-treat principle.

Algorithmic Confounding

To assess the robustness of our social-influence estimates to algorithmic confounding, we performed two sensitivity analyses that make use of the particularities of Spotify's recommender system (Figure S7). First, we used the fact that Spotify's key recommendation function, its *Discover Weekly* playlist, is released on Mondays and receives most attention early in the week. For verification, panel A shows the proportion of Twitter mentions of the automated playlist during the observation period January–November 2016.[2] The playlist pushes users to music they have not listened to but that is algorithmically adjacent to the tastes they have expressed on Spotify. If *Discover Weekly* had a confounding influence on our estimates of peer influence, we would expect that removing adoptions that occurred on Mondays or Tuesdays—both in the treatment and in the control group—substantially would reduce the estimated treatment effect, or that it at least would affect our estimates differently compared to removing adoptions from any other weekday. Panel B shows that this is not the case; the social-influence estimate remains stable regardless of the removal of specific weekdays of adoption.

Second, we estimated separate treatment effects depending on the degree that an artist was listed on any chart-like lists generated by Spotify at the time of treatment (panel C). Our data provide weekly information on the inclusion of artists on a total of 3,133 Spotify lists such as "Today's Top Hits," "Pop All Day," or "Hot Country" that were prominent at the time of data collection. We counted the number of lists that included a particular artist in a given week, derived an artist-standardized measure of this count, and interacted the treatment indicator with that standardized count in the artist fixed-effects analysis of Table S3. If our estimates of social influence were confounded by such playlists, one would expect the treatment effect to be higher in periods in which artists are listed to a greater degree. This is not the case. Instead, estimated treatment effects decrease with the number of listings. This finding further supports the result that social influence is more effective for less familiar items.

[2]We collected a total of 101,000 tweets containing the token "discoverweekly," "#discoverweekly," or "discover weekly." Data were collected in April 2021 using the then available Twitter API for academic researchers. We thank Felix Lennert for research assistance.

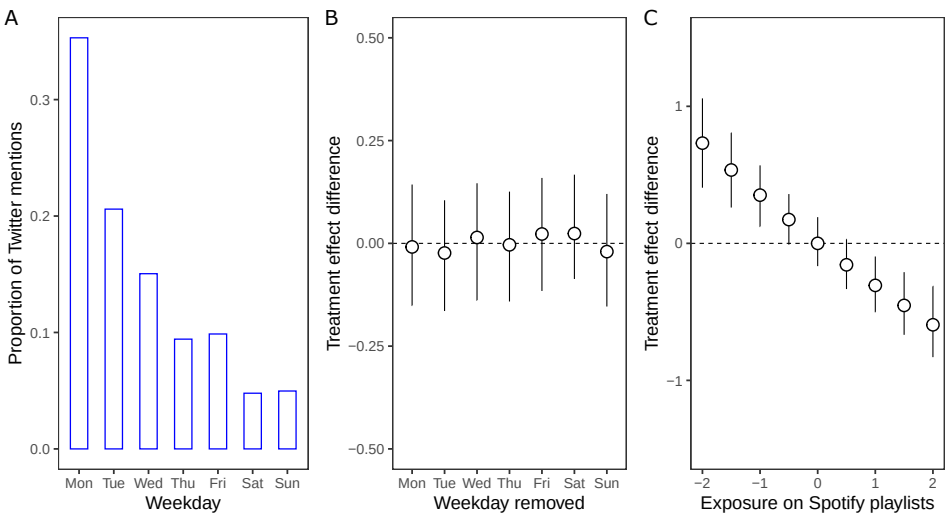

FIG. S7. The social-influence effect is unconfounded by Spotify's recommender system. (*A*) Mentions on Twitter show that *Discover Weekly* receives most attention early in the week (it is released on Mondays). (*B*) Average treatment effects—compared to the population-level treatment effect of 3.6±0.2 (dashed line)—after the removal of adoptions occurring on specific weekdays (in the treatment and in the control group). (*C*) Treatment effect interacted with an artist-standardized measure of the frequency of appearance on Spotfiy's curated playlists (0 corresponds to the mean appearance during the period of study, and 1 a one standard deviation increase in the frequency that an artist feature on Spotify's curated playlists).

S8 Simulation Setup

We built the simulation model in three steps. First, we create populations of agents and artists within genres and assign music tastes to the agents. Second, agents form ties based on their similar tastes. Third, we simulate diffusion processes within the network.

Creating populations

We assume two populations, one of N agents and one of K artists, with the latter being nested in C genres. For the simulation results presented in Figure 4, we use $N = 500$, $K = 100$, and $C = 10$. To create taste profiles of musical preferences, we assume that each agent likes $c < C$ genres. An agent either likes or does not like a given genre such that a C-dimensional vector of 1s and 0s defines an agent's taste profile. Agents' liking of particular genres is assigned through multinomial sampling (see details below). Each genre, in turn, is defined by k artists. Hence, there is a total of $K = k \times C$ artists in the model. To instantiate

popularity bias—that some artists are expected to receive more listening than others, bar any social influence—we let some genres have more listeners, and we assign a non-uniform popularity to artists within genres. More specifically, inequality in genre likes is instantiated by assigning genres to users through multinomial sampling from $b_d \sim \mathcal{N}(100, \sigma_g)$ $/c$, where b_d represents relative popularity of genre d, σ_g controls the degree of inequality across genres, and c is a normalizing constant. Similarly, the relative popularity of artist j in genre d is assigned as $p_{jd} \sim \mathcal{N}(100, \sigma_a)/c$, where σ_a controls the degree of artist inequality within genres and c is, again, a normalizing constant.

Forming networks

To create network ties between agents, we first compute the taste distance between all pairs of agents. We consider the theoretical constructs of "no," "partial," and "complete" taste overlap: If two agents like exactly the same genres (complete overlap), their taste distance equals 0. If they have no preferred genre in common (no overlap), taste distance is 1. If they share preferences for at least one genre but not all (partial overlap), their taste distance is set to 0.5. Having defined taste distance between all agents in the population, we create ties between agents as follows:

1. We select an agent i randomly from the set of all agents whose existing outgoing ties are below a pre-defined outdegree size, Z. Agents start out with no outgoing ties, and in each round, outdegree increases by one for the focal agent.

2. We then randomly draw a taste distance x with probability $P(x) = ce^{-\alpha x}$, where α is a parameter regulating the strength of homophily, and c is a normalizing constant. High (low) values of α imply a strong preference for ties to similar (dissimilar) others. By determining the taste overlap of interconnected agents, α regulates the width of social influence that can occur on the respective network.

3. Among agents with a taste distance x to agent i, we choose an agent j, to whom i creates an outgoing tie with probability $P_{ij} = ce^{-(\omega v_j + \theta q_{ij})}$, where v_j is the current indegree of agent j, q_{ij} is a binary variable indicating whether i and j have shared contacts, and ω and θ are parameters regulating status and triadic closure effects on tie formation. High (low) values of ω produce more (less) unequal indegree distributions. High (low) values of θ produce more (less) clustered network structures.

4. We repeat steps 1–3 until agents, on average, follow the same number of other agents, i.e., have the same preset outdegree size Z.

This procedure overlaps considerably with the homophily-centered network generating models used in Centola (2015) and Zhao and Garip (2021). What distinguishes our procedure is that (1) we do not assume full reciprocity but also consider asymmetric relationships, and (2) we add parameters for regulating the extent of triadic closure and preferential attachment. This enables us to vary the degree of homophily (taste overlap) while keeping constant two central features of social networks, their degree distributions and level of clustering.

Modeling diffusions

To simulate adoption behavior, we use a standard social influence model (e.g., Flache et al. 2017), where adoption behavior is influenced by both the behaviors of others and one's own preferences. More specifically, at any given iteration t, an agent i chooses to listen to artist k with probability $P_{ikt} = s(c\sum_j w_{ij}I_{jkt-1}) + (1-s)T_{ik}$, where $s \in [0,1]$ is the relative weight attached to social influence compared to agent i's own preferences. If $s = 1$, then i's listening behavior is completely determined by what their Alters listened to in the previous iteration. Conversely, if $s = 0$, then agent i chooses what artist to adopt completely independent of others. T_{ik} reflects agent i's preference for song k and is the probability that agent i adopts song k independently. I_{jkt-1} is a binary variable that equals 1 if Alter j adopted song k at time t–1, and 0 otherwise. w_{ij} is a weight that determines how taste distance regulates social influence and equals 1 if i and j share complete taste overlap, 0.5 for partial, and 0 for no overlap, respectively. Finally, c is a normalizing constant. With this adoption function, the simulation proceeds as described in the main text:

1. In the first iteration, each agent adopts one alternative based exclusively on their individual taste.

2. From the second iteration onward, each agent adopts one alternative based partially on their tastes and partially on which alternatives their Alter(s) adopted in the previous iteration, with parameter s determining the relative importance of each.

3. We repeat step 2 until approximate convergence is achieved, which is considered to occur when there is no statistically significant change in unpredictability between consecutive samples of 100 data points, taken 10,000 iterations apart (cf., DellaPosta et al. 2015).

4. Steps 1–3 represent one simulation run. The results reported below are the averages across 30 runs, with 95% confidence intervals.

The simulation results consist of adoption frequencies for each artist, and we examine how the "width" and "strength" of social influence jointly affect the unpredictability of these adoption frequencies. Specifically, we explore how social influence changes the extent to which the rank order of the artists' popularity can be predicted on the basis of the individuals' preferences. To measure unpredictability, or the "unexpectedness" of outcomes, we use the Spearman rank correlation ρ between aggregated individual tastes and aggregate behavior, where 0 (1) indicates no (perfect) correspondence between artists' expected popularity ranks (based on agents' initial preferences) and artists' simulated ranks. In this sense, $\rho = 1$ corresponds to no decoupling and $\rho = 0$ to complete decoupling. Our argument is that social influence must be both strong and wide in order to bring about unexpected outcomes. If social influence is strong but narrow, the rank order of artists will remain intact.

S9 Simulation Robustness Checks

The simulation results presented in the main text assume a particular configuration of the auxiliary parameters of the simulation model. These parameters determine various properties of the population (the number of agents, topics, and songs, and the level of popularity bias), and various properties of the network (the tendency for triadic closure and preferential attachment). In this section, we demonstrate that the qualitative nature of our results are not dependent on these choices.

Extending upon our simulation analysis, future work could incorporate further mechanisms of diffusion, such as complex contagion, as well as investigate higher-order network dynamics in greater detail, identifying the structural conditions under which a minimal set of partially overlapping ties can still enable effective decoupling. In addition, future work could build on these simulations to explore the interaction between social and algorithmic influence in bringing about the emergence of the unexpected.

Figure S8 replicates the main simulation results based on networks that exhibit less clustering (panel *A*) and less indegree inequality (panel *B*), respectively, than those in the main text. While the main simulation results are based on networks with clustering corresponding to a transitivity score of 0.3 and indegree inequality corresponding to a Gini coefficient of 0.4, we use a transitivity score of 0.1 and Gini 0.2, respectively.

In the main simulation analysis, we assumed that the users' music tastes would remain constant over time. Users are influenced by others, but do not change in what genres they like. We relax this assumption in Figure S9, where music tastes are allowed to evolve. More specifically, we introduce an additional parameter, τ, that determines the probability

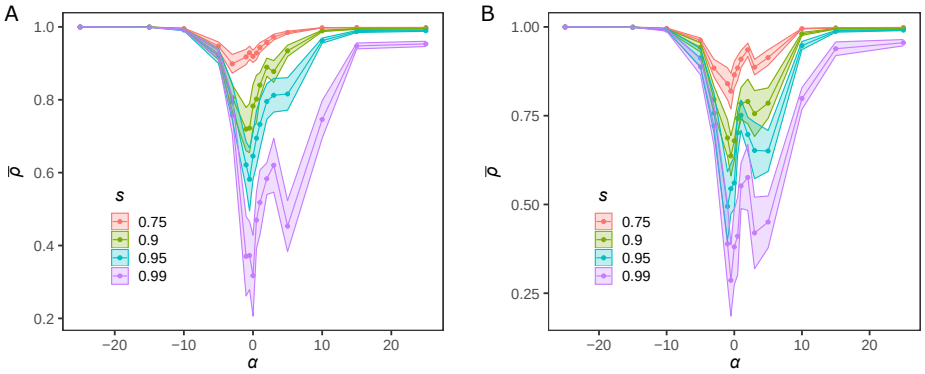

FIG. S8. Replication of the main simulation result using alternative parameter settings for the underlying networks. (*A*) Lower clustering (transitivity ≈ 0.1 instead of 0.3). (*B*) Lower indegree inequality (Gini ≈ 0.2 instead of 0.4). As in the main Figure 4*A*, the panels depict the degree of decoupling (ρ) between artist popularity and aggregate tastes under varying levels of strength (s) and width (α) of social influence. $\rho = 1$ indicates no decoupling, while $\rho = 0$ indicates complete decoupling. Each point in the plots reflects the average across 20 runs ($\bar{\rho}$), and the confidence intervals a 95% confidence interval.

that the focal user expands their taste to include a new genre that they were exposed to.[3] Because the number of iterations in the simulations are on the order of tens of thousands, we consider values of $\tau \in \{0.01, 0.001\}$, instantiating slow taste change. Qualitatively, the main result remains unchanged under different degrees of taste change.

Figure S10, finally, replicates the main simulation results using greater levels of popularity bias (panel *A*) and larger population sizes (panel *B*). In the main simulation, results are based on populations with 500 agents and an artist popularity space with $\sigma_a = \sigma_g = 10$. Here, we increase the number of agents to 1000 and 2000, as well as increase the artist popularity inequality by 50% and 100%. As expected, the absolute level of decoupling decreases as one increases the popularity bias, but the overall shape remains qualitatively similar across levels of artist popularity inequality. Further, population size does not change the main result.

[3]To prevent users from ending up liking all genres with enough iterations, we impose a limit of liking four genres. A user who likes four genres may still update their taste; they then replace one of their liked genres.

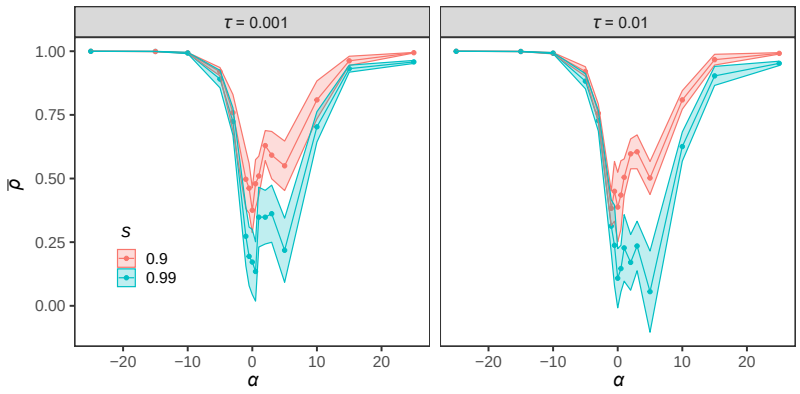

FIG. S9. Replication of the main simulation result allowing agents to change their tastes over time. The main result is not affected by introducing τ and by setting it to different values (0.001, 0.01) which determine the pace of taste change.

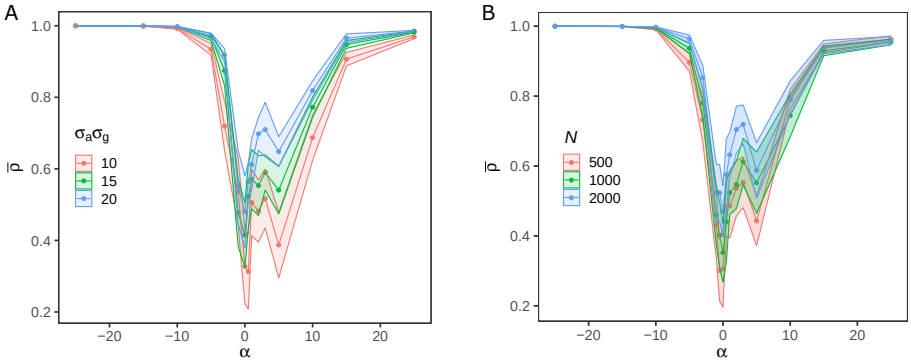

FIG. S10. Replication of the main simulation result (focusing on $s = 0.99$) (*A*) assuming different numbers of agents N and (*B*) different levels of popularity bias σ_a, σ_g. The main result is not qualitatively affected by increasing population size or popularity bias.

Supplementary References

Anderson, Ashton, Lucas Maystre, Ian Anderson, Rishabh Mehrotra, and Mounia Lalmas. 2020. "Algorithmic Effects on the Diversity of Consumption on Spotify." *Proceedings of The Web Conference* 2020:2155–65.

Austin, Peter C. 2009. "Balance Diagnostics for Comparing the Distribution of Baseline Covariates Between Treatment Groups in Propensity-Score Matched Samples." *Statistics in Medicine* 28:3083–107.

Belloni, Alexandre, Victor Chernozhukov, and Christian Hansen. 2014. "High-Dimensional Methods and Inference on Structural and Treatment Effects." *Journal of Economic Perspectives* 28:29–50.

Davison, Anthony C., and David V. Hinkley. 1997. *Bootstrap Methods and their Application*. New York: Cambridge University Press.

Easley, David, Eleonora Patacchini, and Christopher Rojas. 2020. "Multidimensional Diffusion Processes in Dynamic Online Networks." *PLOS ONE* 15:e0228421.

Fleder, Daniel and Kartik Hosanagar. 2009. "Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity." *Management Science* 55(5):697–712.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2011. "MatchIt: Nonparametric Preprocessing for Parametric Causal Inference." *Journal of Statistical Software* 42.

Kuroki, Manabu, and Judea Pearl. 2014. "Measurement Bias and Effect Restoration in Causal Inference." *Biometrika* 101:423-37.

Louizos, Christos, Uri Shalit, Joris Mooij, David Sontag, Richard Zemel, and Max Welling. 2017. "Causal Effect Inference with Deep Latent-Variable Models." *Advances in Neural Information Processing Systems* 30 :6446-56.

McCallum, Andrew Kachites. 2002. "MALLET: A Machine Learning for Language Toolkit." `http://www.cs.umass.edu/~mccallum/mallet`.

Morgan, Stephen L., and Christopher Winship. 2014. *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, 2nd ed. New York: Cambridge University Press.

Park, Patrick S., Joshua E. Blumenstock, and Michael W. Macy. 2018. "The Strength of Long-Range Ties in Population-Scale Social Networks." *Science* 362:1410–3.

Pearl, Judea. 2009. *Causality: Models, Reasoning and Inference*. New York: Cambridge University Press.

Peña, Jose M. 2020. "On the Monotonicity of a Nondifferentially Mismeasured Binary Confounder." arXiv:2005.13245.

Stuart, Elizabeth A. 2010. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science* 25:1–21.

Stuart, Elizabeth A., and Donald B. Rubin. 2008. "Best Practices in Quasi-Experimental Designs." In *Best Practices in Quantitative Social Science*, edited by Jason Osborne. Thousand Oaks, CA: Sage.

Watts, Duncan, and Steven H. Strogatz. 1998. "Collective Dynamics of "Small-World" Networks." *Nature* 393:440–2.