The wisdom of the few: Predicting collective success from individual behavior

Manuel S. Mariani,[1,2,a)] Yanina Gimenez,[3] Jorge Brea,[3] Martin Minnoni,[3] René Algesheimer,[2] and Claudio J. Tessone[2]

[1)] *Institute of Fundamental and Frontier Sciences, University of Electronic Science and Technology of China, Chengdu 610054, PR China*

[2)] *URPP Social Networks, University of Zurich, CH-8050 Zurich, Switzerland*

[3)] *Core Data Science, Grandata, 550 15th Street, San Francisco, California 94103, USA*

Can we predict the future success of a product, service, or business by monitoring the behavior of a small set of individuals? A positive answer would have important implications for the science of success and managerial practices, yet recent works have supported diametrically opposite answers. To resolve this tension, we address this question in a unique, large-scale dataset that combines individuals' purchasing history with their social and mobility traits across an entire nation. Surprisingly, we find that the purchasing history alone enables the detection of small sets of "discoverers" whose early purchases consistently predict success. In contrast with the assumptions by most existing studies on word-of-mouth processes, the social hubs selected by network centrality are not consistently predictive of success. Our approach to detect key individuals has promise for applications in other research areas including science of science, technological forecasting, and behavioral finance.

Whenever we engage in daily activities, we leave traces of our actions. When we purchase in stores using a credit card, our payments are recorded in the credit card's records. When calling our friends, several data (including our communication activity and our geographical location) are recorded in detailed logs by the telecommunication provider. Our increasing ability to collect and analyze human-generated data has fueled interdisciplinary efforts to quantify and predict socioeconomic and behavioral patterns at both the individual and the collective scale[1,2]. A recent stream of studies – which can be grouped under the umbrella term *science of success*[3] – have aimed to uncover common success patterns in systems as diverse as science[4,5], art[6], show business[7], and online social systems[8–11], among others. Mostly rooted in complexity science, these studies have interpreted success as a collective phenomenon that results from a myriad of individuals' actions, and they have examined success patterns of cultural products[12–14], individuals' careers[7,15,16], and teams[17].

Although these efforts have deepened our understanding of the collective dynamics of success, linking individual-level behavioral patterns and collective success remains elusive: In a complex socioeconomic system, are there key individuals whose actions are predictive of collective success? If rich individual-level data are available, how to best detect these individuals? If these key individuals exist and their behavior is persistent over time, organizations or policymakers might track them to make reliable success prediction of businesses, retailers, and products. However, the answers to these questions elude us because of two main challenges. First, it is usually hard to connect various sources of complex individual-level behavioral data and economic success data; many organizations and researchers usually have access to only one of these two kinds of data. Second, some studies attempted to link the behavior of social hubs (i.e., individuals with a disproportionate number of contacts) and the collective success of empirical or simulated diffusion processes in online social networks, but they led to opposite conclusions[8,18–22].

Here, we overcome these challenges by rigorously measuring the predictive power of different sets of key individuals embedded in a large-scale socioeconomic system, and evaluating the predictive power of sets of individuals detected from three different kinds of individual-level behavioral data: purchasing, social, and mobility data. More specifically, we analyze an anonymized dataset[1] that includes individuals' time-stamped purchases in brick-and-mortar stores in an emerging country in America, and their social network based on mobile phone communication [1]. We extract individuals' purchases from a large-scale Credit Card Record (CCR) provided by a major bank (from June 2015 to May 2018); we extract individuals' communication and mobility patterns from a massive Call Data Record (CDR) provided by a mobile phone operator (from January to December 2016) operating in the same country as the bank. Importantly, we can match a substantial number of individuals[2] in the CCR with the individuals in the CDR (see Supplementary Note S1). Differently from recent works on success prediction in science[5] and online social systems[11], our unique anonymized dataset[1] allows us to study the success prediction problem for (nameless) physical stores

[a)] Electronic mail: manuel.mariani@business.uzh.ch

[1] All the data analyzed in this article are anonymized. The subjects of the analysis (individuals and stores) are represented by meaningless hashes in the dataset. All individuals are non-identifiable, all stores are nameless, and all transactions are innominate.

[2] Each individual in our dataset is *non-identifiable*, meaning that (s)he is represented by a meaningless hash in the dataset, and there is no way to reconstruct the individuals' real identities.

Category	T	S_{tot}	S_{tr}	S_{val}	$\langle c \rangle$	I	$\langle k \rangle$
Eating places	9,946,197	115,383	28,196	11,654	154.44	1,323,647	7.51
Food stores	13,606,954	97,096	15,046	6,456	240.26	2,061,374	6.60
Clothing stores	3,967,272	66,753	12,809	3,822	92.80	1,166,701	3.40

TABLE I. **Basic data properties.** We report basic properties of the (nameless) eating places, food stores, and clothing stores found in the Credit Card Record (see Supplementary Note S1). The variables represent: the total number of first-time transactions in stores, T, over the training period; the total number of stores that received transactions, S_{tot}; the number of stores that received their first transaction within the training period (excluding the last two months), S_{tr}; the number of stores that received their first transaction within the validation period (excluding the last two months), S_{val}; the average number of first-time customers per store, $\langle c \rangle$, over the training period; the number of individuals who made purchases during the training period (excluding the last two months), I; the average number of stores visited by the individual during the training period (excluding the last two months), $\langle k \rangle$.

at the national level, and to connect it with the individuals' purchasing, communication, and mobility patterns.

We find that the early purchases by social hubs (detected as the most central nodes in the social network built from the Call Data Record) are not consistently predictive of success for the stores where they purchase. Motivated by this finding, differently from the vast body of literature on word-of-mouth diffusion and influential nodes in complex networks[22–27], we search for key individuals[2] directly from the credit-card transaction time series, without using social-network information. Our analysis of the Credit Card Records reveals the existence of discoverers[28]: a small set of individuals[2] who repeatedly purchase in recently-opened stores that later become successful. We find that the early purchases (or lack thereof) by the discoverers are highly predictive about the eventual success of the stores where they purchase. The discoverers' purchases have a stronger predictive power than those by social hubs[8,25], high-expenditure individuals, and explorative individuals detected through mobility-related traits[29,30] and their number of visited stores[3].

While there is little overlap between the groups of discoverers for different store categories, the discoverers are characterized by socioeconomic traits that are consistent across categories. In particular, they exhibit above-median total expenditures, number of visited stores, network centrality (inferred from the analyzed mobile phone communication data), and mobility diversity. However, compared to the social hubs, they exhibit a significantly smaller number of contacts, and a larger mobility diversity and number of visited stores, suggesting that they play a substantially different role in the success dynamics. We also observe age and gender differences for the discoverers of different categories of stores.

Our findings support the *wisdom of the few* paradigm: in order to predict the future success of a recent agent in a large socioeconomic system, one can rely on the actions by a small set of individuals. The detection of this set only requires the transaction history, enabling researchers and organizations to early predict the success of a business without the need for social network data nor mobility data. The discoverer detection procedure and the predictive scheme developed here are general and can be applied for success prediction in other complex systems like scientific and technological innovation, and financial markets.

[3] For the sake of convenience, we refer to a store where an individual made a purchase as a "visited store". We have no access to physical visits to stores that did not result in a purchase.

RESULTS

Credit card record and call data record

We analyze a unique anonymized dataset[1] that includes credit-card transactions of (non-identifiable) individuals[2] over a three-year temporal window (from June 2015 to May 2018). By recording only the first purchase of each individual[2] in each nameless store and after pre-filtering the data (see Supplementary Note S1), the complete dataset includes more than 140 million of time-stamped transactions[1]. In the following, whenever we will refer to an individual, this should be interpreted as a non-identifiable individual whose real identity is impossible to identify. Accordingly, whenever we will refer to a group of "detected" individuals, this should be interpreted as a group of individuals whose real identities cannot be identified. For the sake of better readability, when referring to the analyzed datasets and the individuals in the following, we shall omit the labels "anonymized" and "non-identifiable".

The first observation is that the dataset is highly heterogeneous, encompassing categories as diverse as book stores, tech stores, and florists, among many others. To control for this heterogeneity, we restrict our analysis to three categories of stores that have a well-defined interpretation: eating places, clothing stores, and food stores. The stores that belong to each of the three categories are selected based on their Merchant Category Code (MCC) information present in our database – we refer to Supplementary Table S1 for details.

We split the dataset's time span into three periods:

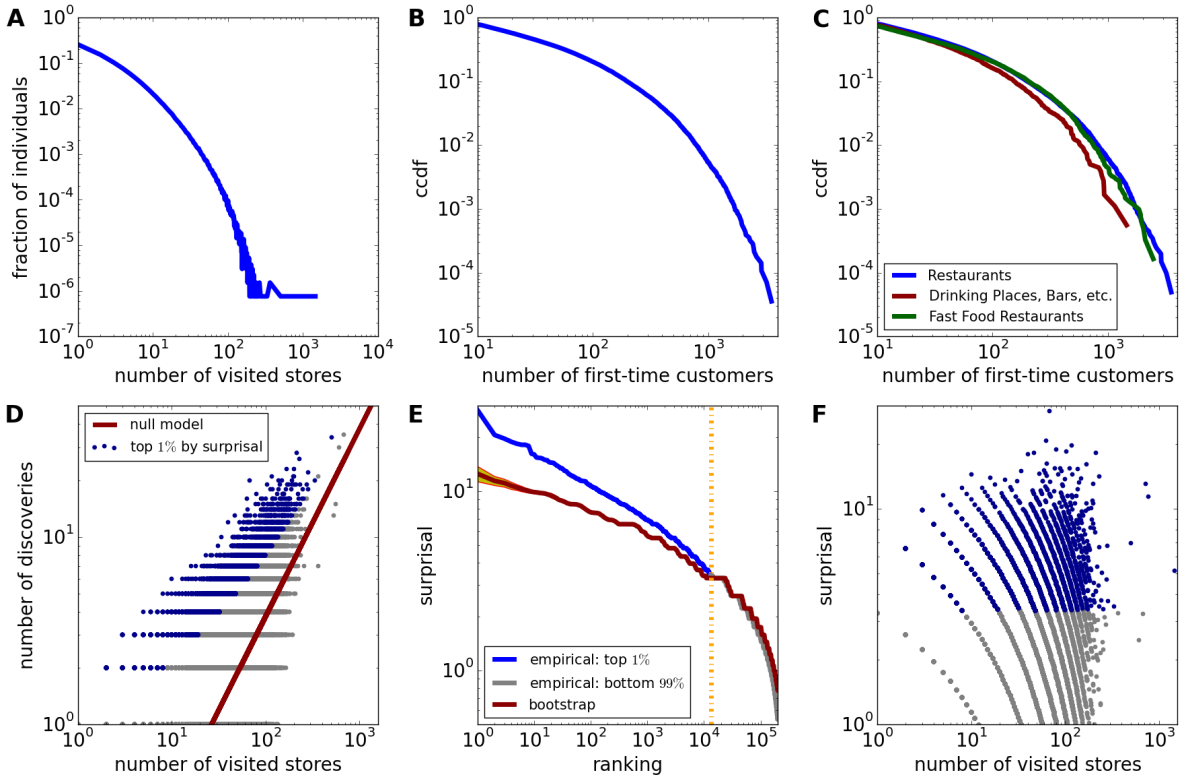

FIG. 1. **Individuals' activity and discovery patterns for eating places. (A)** Distribution of the number of visited stores per individual. **(B)** Complementary cumulative distribution function of the number of first-time customers per store. **(C)** Same as (B) for each Merchant Category Code (see Table S1). **(D)** Number of discoveries vs. number of visited stores. The two variables are positively correlated, yet the trend deviates from the expected one under the null model, and some individuals collect significantly more discoveries than expected from their activity alone. The straight line represents the expectation under the null model, whereas the blue dots mark the top-1% individuals by surprisal. **(E)** Zipf plot of the individuals' surprisal for the empirical data and the null model. The discrepancy between the two curves is significant, meaning that the surprisal values of most of the top-1% individuals by empirical surprisal (i.e., those at the left of the vertical dashed line) are significantly above the surprisal values expected for their ranking position. **(F)** Surprisal vs. number of visited stores. The two variables are only weakly correlated.

a 6-month pre-filtering period that is used to determine which stores appear for the first time in the training period that follows; a 18-month training period that ranges from December 2015 to May 2017, where we aim to detect groups of key individuals (described below); a 12-month validation period that is used to perform and validate success predictions. The main rationale behind our choice of the relative duration of training and validation period is that the validation period should be long enough to include a substantial number of new stores to validate our predictions, yet short enough to not exceedingly restrict the training period where we detect the key individuals.

Table I summarizes basic data properties. We denote as k_i and c_α the number of stores visited by individual i and the number of first-time customers of store α, respectively. The individuals' number of visited stores, k, is highly heterogeneous, with a small number of outliers with a large number of visited stores (Fig. 1A and Figs. S1-S2). Similarly, the number of first-time customers per store is highly heterogeneous (Figs. 1B) and dependent on the store category (Figs. 1C and Figs. S3-S4). The number of new stores per month (i.e., stores that receive their first transaction) shows a decreasing trend over time (Fig. S5). Besides the CCR, we analyze a CDR from the same market over a one-year period that overlaps with the CCR's time span. For a subset of the population, we know both the social behavior (in terms of mobile phone communication) and the economic behavior (in terms of monetary purchases in stores) – see Supplementary Note S2 for details. From the CDR, we extract individual-level traits related to their centrality in the social network (time-averaged degree[8], collective influence[25], and social diversity[31]), and mobility (radius of gyration[29] and mobility diversity[30]). We refer to Supplementary Note S3 for the details of the measured features.

Framing a success prediction problem

We consider the classification problem where we aim to predict whether a store introduced in the validation period will be among the top-10% shops by final number of first-time customers, among the stores with the same MCC that received their first transaction in the same month – if this is the case, we say that the store is *successful*. Our definition of success factors out two potential confounding factors in our measure of success: store age and category (see Fig. S4). To quantify the predictive power of a given group of individuals, $\mathcal{I}$, we measure the fraction of successful stores among those that featured an individual in $\mathcal{I}$ among the earliest w first-time customers – we refer to this fraction as $\mathcal{I}$'s *success rate*. We divide this success rate by the baseline success rate given by the fraction of successful stores among those that received at least w first-time customers by the end of the validation period, obtaining the fold increase of $\mathcal{I}$'s success rate with respect to the baseline expectation. For the sake of brevity, we refer to this ratio as $\mathcal{I}$'s *success-rate fold increase*. For each group $\mathcal{I}$ of individuals, we also measure alternative prediction evaluation metrics (recall, positive likelihood ratio, and Matthews' correlation coefficient) – see Supplementary Note S4 for details.

Importantly, the above-defined problem allows us to quantify and compare the predictive power of the actions by different groups of individuals, offering the opportunity to clarify whether some groups of individuals exhibit a strong and consistent predictive power. The problem is fundamentally different from the influence maximization problem that has focused on the detection of vital nodes for the structure of and dynamics on a given network[22,25,26]. Based on that stream of literature, it is natural to investigate whether sets of individuals that perform well in influence maximization problems (e.g., selected by degree[8] or collective influence[25]) make purchases that are predictive of success. The intuition would be the following: if social hubs have a disproportionate impact on diffusion processes, their purchases in recent stores might trigger large-scale word-of-mouth processes and, as a result, be predictive of success for the visited store. However, we find that this is not always the case.

The inconsistent predictive signal from the hubs' purchases

We find that the social hubs (selected by appropriate time-averages of degree[8] and collective influence[25], see Supplementary Note S3 for details) are not reliable predictors of success for the visited stores. For example, for eating places that received their first transaction within the validation period, the hubs' success rate is 13.63% by considering the earliest 10 first-time customers, against the 11.38% baseline success rate, resulting in a 1.20-fold increase in success rate. The success-rate fold increase is even larger for social hubs selected by collective influence (1.27). These results indicate that the social hubs' purchases are predictive of success for eating places and that, remarkably, the top-individuals by collective influence are better predictors of success than the top-individuals by degree. We observe a qualitatively similar result for food stores, but not for clothing stores: When considering the earliest 10 customers of a clothing store, the stores that received a purchase by a social hub (selected by degree) exhibit a success rate that is only marginally larger than the baseline (12.85% against 12.80%). The hubs' success rate can be even smaller than the baseline for larger numbers of early customers included to make the prediction (see Fig. 2 and the related discussion below), and similar results are obtained with other prediction evaluation metrics (see Fig. S6). We conclude that in general, the early-purchases by the hubs are not reliable predictors of success for the visited store. The observed inconsistent performance of the social hubs aligns with the system-dependent conclusions on the role of social hubs for success prediction in online systems[21]. Additional research is needed to uncover the reason why the social hubs are only predictive of success for specific store categories, which might reflect the different importance of word-of-mouth processes for the success of different categories of stores.

Discoverer detection

Motivated by our predictive problem and the inconsistent predictive signal from the social hubs, we adopt a statistical procedure that seeks to find individuals – referred to as *discoverers*[28] – who are persistently able to discover stores with a high potential of becoming popular. To this end, we define a *discovery* as an early purchase in a store that later becomes successful. The discoverers are selected by a measure of statistical unexpectedness – called *surprisal*, a term coined in thermodynamics[32] – that quantifies how unlikely the observed number of discoveries was under a null model that preserves the individuals' level of activity (in terms of number of visited stores) and assumes that everyone has the same likelihood to collect a discovery – see for the details. In general, it is not guaranteed that there exists a sizeable set of individuals whose number of discoveries significantly deviate from the expectations from the null model, once we account for the possibility of achieving an unexpected number of discoveries by statistical fluctuations.

However, we find that while the number of discoveries per individual is positively correlated with the number of visited stores (Fig. 1D), the deviations from the trend are significant and cannot be explained by chance. To rule out the possibility that high values of surprisal are obtained through random fluctuations, we compare the empirical surprisal values of the detected discoverers with the top-surprisal values observed by resampling the individuals' number of discoveries from the null-model distribution (see Material and). We find that the largest empirical surprisal values are significantly larger than the

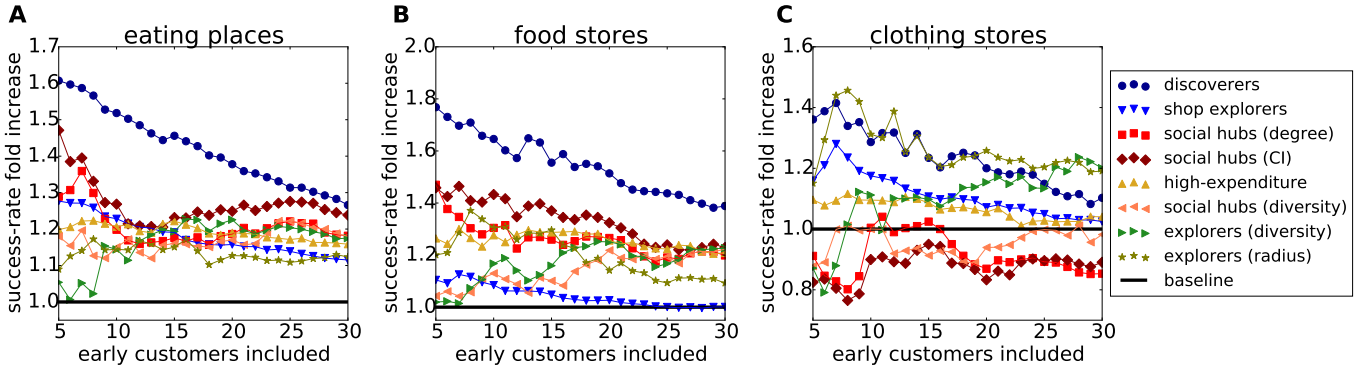

FIG. 2. **Wisdom of the few: Success predictions based on groups of key individuals.** Success-rate fold increase for stores visited early on by different groups of top-individuals, for eating places **(A)**, food stores **(B)**, and clothing stores **(C)** that received their first transaction within the validation period. The discoverers' success rate is substantially larger than that of other groups for eating and food stores. The discoverers and the explorers selected by radius of gyration are similarly competitive for clothing stores, and they outperform the other groups. Compared to the discoverers, the performance by other groups of selected individuals is inconsistent or weaker. For example, the social hubs are associated with an above-baseline success rate for eating places and food stores, but not for clothing stores. The explorers (selected by radius of gyration) exhibit a success rate comparable to the discoverers' one for clothing stores, but their predictive power is substantially weaker for eating places and food stores.

largest surprisal values obtained by resampling the individuals' number of discoveries (Fig. 1E). The surprisal values are little correlated with the individuals' number of visited stores, indicating that activity alone is not a good proxy for the propensity of an individual to discover successful stores (Fig. 1F). The results in Fig. 1 refer to eating places; results for other store categories are qualitatively similar (Figs. S1-S2). Taken together, these findings indicate that some individuals exhibit a clear propensity to purchase in recently-opened stores that later become successful, and it is highly unlikely that this pattern can be explained solely by the individuals' level of activity.

Tracking detected individuals to predict store success

The previous analysis reveals that there exist individuals who repeatedly purchase in recent stores that later end up being successful. Yet, the detected individuals would have true predictive value only if they would be predictive of success over a time window that follows the training period. Is the discoverers' tendency to collect discoveries persistent enough over time to allow us to track their actions for reliable out-of-sample success predictions? This is not obvious *a priori*, but if it would be the case, it would suggest that discovering successful stores is a persistent behavioral trait of the discoverers. Besides, are the predictions made by tracking the discoverers' purchases more accurate than those obtained through other groups of top individuals, selected by network centrality measures (*hubs* selected by degree[8], collective influence[25], and social diversity[31]), total expenditures (*high-expenditure* individuals[33]), number of visited stores (*store explorers*), and mobility-related features (*explorers* selected by mobility diversity[30] and radius of gyration[29])? We refer to the Supplementary Notes S2–S3 for all details on the detection of the groups of top-individuals included in the analysis. Driven by these questions, we study the predictive problem framed above, where the predictive power of top-individuals detected within the training period is evaluated for stores that received their first transaction within the validation period.

We find that the detected discoverers exhibit a consistent predictive signal across all three store categories (Fig. 2). In particular, for eating places and food stores, the discoverers exhibit the largest success-rate fold increase for all numbers of included early customers (Figs. 2A–B). For clothing stores, explorers selected by radius of gyration exhibit a similar success rate to the discoverers' one (Figs. 2C). The early purchases by other classes of individuals might still be associated with larger-than-baseline success rates, yet none of the other groups of individuals is competitive across all three store categories. This is true not only for the social hubs (see the discussion above), but also for high-expenditure individuals and explorers – we refer to Fig. 2 for the full results.

Similar conclusions can be drawn from the results obtained with two more prediction evaluation metrics: the Matthews' correlation coefficient and the positive likelihood ratio[34] – we refer to Fig. S6 for the detailed results. The recall metric (namely, the fraction of successful stores that are classified as successful) exhibits a different trend compared to the success rate because by construction, it favors groups of individuals that purchased in many different stores. Because of this, store explorers exhibit the largest recall values across all three categories, followed by the discoverers and high-expenditure individuals (see Fig. S6). Our results are reasonably ro-

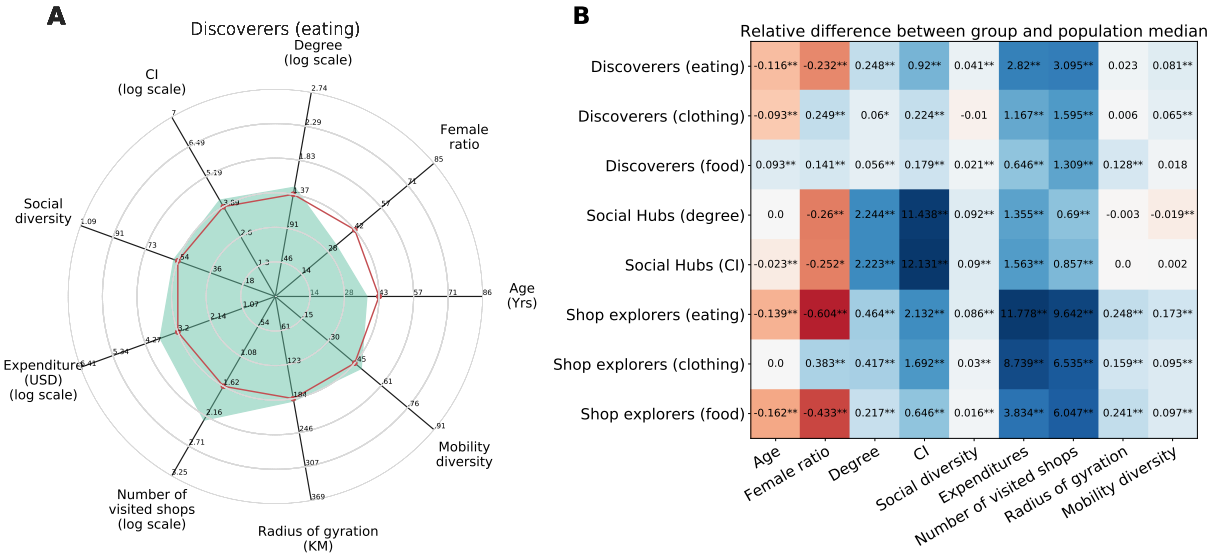

FIG. 3. **Profiling key individuals: Discoverers, social hubs, and explorers. (A)** The radar plot illustrates the profiling of a group of individuals (discoverers of eating places). We compare the individuals' traits (represented by the radars' axes) with the traits' population medians (connected by the red polygon). **(B)** For each group of individuals (row), $\mathcal{I}$, and each trait (column), $\mathcal{T}$, we report the difference between $\mathcal{T}$'s median for individuals in $\mathcal{I}$ and the population median, normalized by the population median. Some of the discoverers' socioeconomic and mobility traits are consistent: regardless of store category, the typical discoverer exhibits above-median total expenditures, number of visited stores, collective influence, number of social contacts, and mobility diversity. By contrast, the discoverers' demographic traits (age and gender) are category-dependent. The traits of social hubs and store explorers are also reported. Alongside the reported relative differences, single (*) or double (**) stars mark significance levels below $P < 0.05$ and $P < 0.01$, respectively, under the binomial test for female ratio, and the Mood's median test for all other features (see Material and).

bust with respect to variations in the parameters of the analysis – see Supplementary Note S5 and Figs. S8–S14 for the detailed results. We also notice that it is possible to consistently improve the predictions' success rates by pairing the discoverers with other groups of individuals, yet these improvements tend to be marginal (see Supplementary Note S5 and Fig. S7).

Beyond the classification problem, one can also investigate whether stores that received a discoverer among the earliest customers tend to receive a larger number of customers in the future. We find that this is the case across all three store categories. For example, among the eating places that received their first transaction in the validation period, those that had a discoverer among the earliest 10 customers tend to gain approximately 40.3% more first-time customers that the average store with the same age and category. Similar results are obtained for food and clothing stores, where the relative fold increases of the final number of customers associated with the presence of a discoverer among the earliest customers are 33.2% and 17.0%, respectively (see Fig. S15 for the detailed results as a function of the number of early customers included in the analysis). Taken together, our results indicate that early purchases by the discoverers are associated with increased odds of success and an increased number of customers with respect to the baseline.

Socioeconomic characterization of discoverers, social hubs, and explorers

Having detected the discoverers and measured their predictive power, it is inevitable to investigate which traits make them different from ordinary individuals and social hubs. This is relevant for companies in order to detect prospective discoverers in scenarios where transaction data for their customers are not available, and potentially nudge into behaving as discoverers individuals who exhibit appropriate combinations of traits. We start by observing that there is little overlap between pairs of groups of discoverers of different categories of stores: The largest overlap is 380 common discoverers among the discoverers of eating places and food stores (corresponding to the 2.8% of the discoverers of eating places and the 1.8% of the discoverers of food stores); the largest Jaccard similarity is 0.015 observed between the discoverers of eating places and clothing stores. This indicates that it is unlikely that a discoverer in one category is also a discoverer in another category of stores, suggesting that the discoverers are highly specialized.

Interesting demographic differences emerge across the groups of individuals (Fig. 3). The discoverers' demographic traits are consistent with the store explorers' ones for eating places (the members of both groups tend to be males whose age is below the population median) and for clothing stores (the members of both groups tend to be

female). A clear difference emerges for food stores: the store explorers tend to be males with below-median age, yet the discoverers tend to be female with above-median age. Hence, not only store explorers are not necessarily discoverers (Fig. 1F), but the two groups of individuals can also exhibit starkly different demographic traits.

As for socioeconomic traits, across all three store categories, the trait where the discoverers' median deviates the most from the population median is the number of visited stores, followed by total expenditures, collective influence, and number of contacts (for all these differences, $P < 0.01$ according to the Mood's median test). The discoverers' large expenditures and number of social contacts suggest that they have a higher socioeconomic status and degree of social connectedness than ordinary individuals. In this sense, they respect the traits of early adopters outlined in Rogers' seminal work on the diffusion of innovations[35]. Yet, the discoverers are not outstanding in any of these traits: for example, the social hubs selected by degree exhibit a substantially larger number of contacts, and store explorers purchase in a substantially larger number of stores. We also notice that the discoverers' mobility diversity is slightly above the population median, with the smallest relative difference observed for the discoverers of food stores ($+2.0\%, P < 0.05$). Overall, these results support a scenario where the discoverers benefit from various socioeconomic traits, yet none of the investigated traits is predictive of the behavior of an individual as a discoverer. Further research is necessary to assess whether alternative socioeconomic traits, or some psychological traits, can be leveraged to accurately distinguish the discoverers from the rest of the population.

Intriguingly, we can compare the socioeconomic and demographic traits of the social hubs with results obtained by previous studies[2]. We find that the social hubs (selected by degree and collective influence) exhibit above-median expenditures ($P < 0.01$). This is in qualitative agreement with previous studies that reported that expenditures are a reliable proxy for monthly income[33], and social hubs tend to have a higher economic status[36]. The social hubs also exhibit an above-median number of visited stores, but not an above-median radius of gyration nor mobility diversity. In line with previous findings on online platforms[8], social hubs tend to be males. Compared to the discoverers, the social hubs exhibit a smaller radius of gyration and a smaller number of visited stores. Store explorers tend to be more central in the social network and to travel for a longer distance (as measured by their radius of gyration) than the discoverers.

DISCUSSION

This article focused on a data-rich scenario where we have access to individual-level purchasing history, communication activity, and mobility information. Our results support the *wisdom of the few* paradigm: only a small set of individuals act as reliable success predictors, and we can accurately detect them from the transaction history through a parsimonious and highly-scalable statistical procedure, without the need to integrate social and mobility information. This is highly non-trivial because of the several factors that can limit success predictability in complex socioeconomic systems[11,37], and the recent contrasting conclusions on the possibility to predict success based on key individuals[8,18,19,21,38]. Our general approach to measure the predictive power of individuals has promise for applications in different social systems, and additional research is needed to establish the degree of universality of our results.

Most of the influencer-related literature aims to detect influential individuals from their position in a social network[22,26,27]. However, our results reveal that early purchases by social hubs are only predictive of success for specific store categories, and their predictive signal is weaker than that observed for discoverers. By contrast, the discoverers are not among the most central individuals in the social network that we analyzed, yet their monetary transactions are predictive of the stores' potential, without the need to explicitly consider the social network the discoverers are embedded in. Therefore, companies and organizations that have only access to transaction data would be able to use our approach to detect key individuals for success prediction.

This paper extends the recent stream of literature on success prediction[4,7,11,13] by focusing on brick-and-mortar stores. In today's world where online purchasing is disrupting traditional offline retailing, organizations, marketers and investors can benefit from understanding the factors that can drive a store to long-term success. This is relevant not only to traditional brick-and-mortar stores, but also to companies that aim to combine online and offline retailing activities to improve sales and customer experience[39,40]. In line with the growing number of studies on popularity prediction in online systems[11] and impact prediction in science[4], our definition of store success builds on a popularity-based metric: the number of first-time customers received by a store[39]. This choice allowed us to ask a well-defined predictive question. On the other hand, alternative metrics (e.g., the number of regular or satisfied customers) may capture different and equally relevant dimensions of store success. Understanding how to best quantify and predict store success is a challenge for future research, and it may require a combination of popularity-based and monetary indicators.

To conclude, three limitations of our study need to be addressed in future studies. First, our study implicitly assumed that mobile-phone communication data provide us with sufficiently good estimates of the individuals' centrality in the actual network of social contacts. On the other hand, with the rise of online social networks and instant messaging platforms, the phone communication network only provides us with a partial representation of the actual communication flows in society. Obtaining a complete representation of the social communication

patterns across an entire nation is clearly unattainable. Therefore, while the incompleteness of our social graph is a limitation of our work, it also mimics a real-world managerial scenario where an organization has only access to incomplete social information about its customers. Future studies need to generalize our results to online communities or virtual worlds where data on individuals' social connections, communication, and behavior are (nearly) complete.

Second, as for the interpretability-accuracy tradeoff in success predictions[11], our simple classifiers based on the early purchases by selected groups of key individuals have the advantage of having a clear interpretation in terms of individuals' behavioral patterns. At the same time, we did not attempt to maximize the predictive performance. Machine-learning algorithms that incorporate multiple, diverse features might achieve higher precision. In this direction, an exhaustive analysis of the predictability of store success based on many different features, in line with extensive studies carried out for virality prediction in online systems[9,21], is desirable. In future studies, our classifiers based on specific groups of key individuals can be used as baselines for future studies on early-stage success prediction in diverse kinds of socioeconomic systems.

Finally, based on our data, we cannot establish any causal connection between the discoverers' purchases and the success of the stores. We have addressed the predictive question of whether the purchases by selected key individuals are consistent indicators of future success, but we did not address the question of whether the detected discoverers can accelerate the diffusion processes they participate in, similarly as the long-studied opinion leaders and influencers are assumed to do[22,26,27]. Do the discoverers influence their peers through direct communication, or do they choose stores that have a high potential to gain many customers, without actively influencing their future popularity? Are they influential or susceptible members of their social network[41]? What would happen if the discoverers are chosen as seeds of a diffusion process[27], or prevented from adopting a recent innovation they are potentially interested in[42]? Clarifying the role played by the discoverers in empirical diffusion processes through both observational data analysis and field experiments is an intriguing direction for future research.

I. MATERIALS AND METHODS

Discoverer detection

We aim to quantify individuals' tendency to purchase in recent stores that later become successful. To this end, we define a *discovery* as an event where an individual purchases in a store no later than Δ days after the store received its first transaction, and the store turns out to be successful. A store is considered as *successful* if it is among the top-z% stores by final number of first-time customers (at the end of the training period), among the stores with the same Merchant Category Code that received their first transaction in the same month. Stores that received their first transaction within the last two months of the training period are excluded from the analysis; similarly, transactions in that period do not contribute to the individuals' number of visited stores and discoveries, but only to the stores' number of customers. In the following, we denote the individuals who made at least one transaction in a store included in the analysis as $i \in \{1, \ldots, I\}$, where I is the total number of active individuals.

We determine individual i's propensity to discover successful stores as the unexpectedness of i's observed number of discoveries, d_i^*, in terms of the statistical surprisal $s_i(d_i^*) = -\log P_i(d \geq d_i^*)$, where $P_i(d \geq d_i^*)$ represents the probability that individual i collected d_i^* or more discoveries given its total number of visited stores, k_i. More specifically, the probability $P_i(d \geq d_i^*)$ is determined analytically as follows. We consider a process where each individual i draws (without replacement) k_i marbles out of an urn which contains T marbles, D of which are labelled as "discoveries"; T and D match the empirical number of first-time transactions and discoveries, respectively: $T = \sum_i k_i$, and $D = \sum_i d_i^*$. This choice of the null model aims to factor out individuals' level of activity, k_i, from the surprisal measure. The probability that i achieves d discoveries follows the hypergeometric distribution with mean $\langle d_i \rangle = k_i\, D/T$:

$$P_i(d|k_i) = \frac{\binom{D}{d}\binom{T-D}{k_i-d}}{\binom{T}{k_i}}. \tag{1}$$

Individual i's surprisal is given by

$$s_i = -\log P_i(d \geq d_i^*) = -\log\left(\sum_{d=d_i^*}^{k_i} P_i(d|k_i)\right). \tag{2}$$

The discoverers are detected as the top-τ% individuals by s. Results in the main text have been obtained with $z = 10\%$, $\Delta = 90$ days, $\tau = 1\%$; results for different values of z, Δ, τ are reported in Figs. S7–S9.

The surprisal metric naturally rewards individuals whose observed number of discoveries d_i^* far exceeds their expected number of discoveries under the null model. For example, for eating places, the top-individual by surprisal collected 20 discoveries out of 68 eating places where he purchased. His expected number of discoveries under the null model was 2.56, meaning that he collected roughly eight times more discoveries than expected by chance. The probability he achieved 20 discoveries under the null model is 4.7×10^{-13}, resulting in a large surprisal value ($s = 28.4$).

The bootstrap procedure

Even in a random sampling process, some individuals might still achieve a large value of surprisal due to

statistical fluctuations. To ascertain that the largest observed values of surprisal cannot be explained by chance, we perform a bootstrap analysis[28]. In each realization of the bootstrap procedure, for each individual, we extract its number d_i of discoveries from the hypergeometric distribution (1), and compute the surprisal value, s_i, associated with the extracted number of discoveries. For each bootstrap realization, we rank the individuals according to their surprisal, obtaining the Zipf's plot $s(r)$ for that realization. We then average the Zipf plot obtained over different realizations, and compare the resulting Zipf plot with the Zipf plot corresponding to the empirical surprisal values. The results show that for the highest ranking positions, the empirical surprisal values are significantly larger than the bootstrapped ones (Figs. 1E, S1–S2).

Groups of top individuals

We consider eight groups of top individuals, that can be broadly classified into four categories: Discoverers, social hubs, explorers, and high-expenditure individuals. We provide a brief description of the four categories: (1) The *discoverers* are selected by the surprisal defined in Eq. (2); (2) The *social hubs* (or briefly, hubs) are selected by appropriate time-averages of three centrality metrics extracted from the CDR: degree, collective influence, and social diversity; (3) The *explorers* are selected by three metrics: total number of visited stores (extracted from the CCR), time-average of mobility diversity and radius of gyration (extracted from the CCR); (4) The *high-expenditure* individuals selected by their time-averaged total expenditures. For each store category, we selected the top-τ% individuals for each trait as the top-individuals. This choice corresponds allows us to use the same number of top-individuals for all classifiers. All results in the manuscript have been obtained with $\tau\% = 1\%$; results for more selective thresholds are shown in Fig. S10. We refer to the Supplementary Notes S2–S3 for all details on the detection of these groups of individuals.

Statistical analysis

The statistical significance of the differences between group and population median displayed in Fig. 3 have been obtained through the Binomial test and Mood's median test[43] for the female ratio and all the other traits, respectively. The binomial test allows us to test the null hypothesis that males and females are equally likely to occur in the detected group of individuals. The Mood's median test allows us to test whether the group and population median were extracted from two populations with the same median.

Materials and Data Availability

Source code is available on request to the authors. For contractual and privacy reasons, the raw data is not available. Upon request, the authors can provide appropriate documentation for replication, and they might provide samples of the processed data.

Ethics declaration

The Human Subjects Committee of the Faculty of Economics, Business Administration and Information Technology at the University of Zurich has authorized this research on 29 March 2018. In particular, it has reviewed the information regarding the procedures and protocols in our research, and confirmed that they comply with all applicable regulations.

Acknowledgements

This work has been supported by the Science Strength Promotion Program of UESTC and by the URPP Social Networks at the University of Zurich. MSM and CJT acknowledge financial support from the Swiss National Science Foundation (Grant No. 200021-182659). MSM acknowledges financial support from the UESTC professor research start-up (Grant No. ZYGX2018KYQD215).

References

[1] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabási, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., "Computational social science," Science **323**, 721–723 (2009).
[2] J. Gao, Y.-C. Zhang, and T. Zhou, "Computational socioeconomics," Physics Reports (2019).
[3] A.-L. Barabási, The Formula: The Universal Laws of Success (Little Brown and Company, New York, 2018).
[4] A. Clauset, D. B. Larremore, and R. Sinatra, "Data-driven predictions in the science of science," Science **355**, 477–480 (2017).
[5] S. Fortunato, C. T. Bergstrom, K. Börner, J. A. Evans, D. Helbing, S. Milojević, A. M. Petersen, F. Radicchi, R. Sinatra, B. Uzzi, et al., "Science of science," Science **359**, eaao0185 (2018).
[6] S. P. Fraiberger, R. Sinatra, M. Resch, C. Riedl, and A.-L. Barabási, "Quantifying reputation and success in art," Science **362**, 825–829 (2018).
[7] O. E. Williams, L. Lacasa, and V. Latora, "Quantifying and predicting success in show business," Nature communications **10**, 2256 (2019).

[8]J. Goldenberg, S. Han, D. R. Lehmann, and J. W. Hong, "The role of hubs in the adoption process," Journal of Marketing **73**, 1–13 (2009).
[9]J. Cheng, L. Adamic, P. A. Dow, J. M. Kleinberg, and J. Leskovec, "Can cascades be predicted?" in Proceedings of the 23rd international conference on World wide web (ACM, 2014) pp. 925–936.
[10]T. Martin, J. M. Hofman, A. Sharma, A. Anderson, and D. J. Watts, "Exploring limits to prediction in complex social systems," in Proceedings of the 25th International Conference on World Wide Web (International World Wide Web Conferences Steering Committee, 2016) pp. 683–694.
[11]J. M. Hofman, A. Sharma, and D. J. Watts, "Prediction and explanation in social systems," Science **355**, 486–488 (2017).
[12]D. Wang, C. Song, and A.-L. Barabási, "Quantifying long-term scientific impact," Science **342**, 127–132 (2013).
[13]B. Yucesoy, X. Wang, J. Huang, and A.-L. Barabási, "Success in books: a big data approach to bestsellers," EPJ Data Science **7**, 7 (2018).
[14]C. Candia, C. Jara-Figueroa, C. Rodriguez-Sickert, A.-L. Barabási, and C. A. Hidalgo, "The universal decay of collective memory and attention," Nature Human Behaviour **3**, 82 (2019).
[15]R. Sinatra, D. Wang, P. Deville, C. Song, and A.-L. Barabási, "Quantifying the evolution of individual scientific impact," Science **354**, aaf5239 (2016).
[16]Y. Ma and B. Uzzi, "Scientific prize network predicts who pushes the boundaries of science," Proceedings of the National Academy of Sciences **115**, 12608–12615 (2018).
[17]L. Wu, D. Wang, and J. A. Evans, "Large teams develop and small teams disrupt science and technology," Nature **566**, 378 (2019).
[18]D. J. Watts and P. S. Dodds, "Influentials, networks, and public opinion formation," Journal of Consumer Research **34**, 441–458 (2007).
[19]E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (ACM, 2011) pp. 65–74.
[20]B. Libai, E. Muller, and R. Peres, "Decomposing the value of word-of-mouth seeding programs: Acceleration versus expansion," Journal of Marketing Research **50**, 161–176 (2013).
[21]B. Shulman, A. Sharma, and D. Cosley, "Predictability of popularity: Gaps between prediction and understanding," in Tenth International AAAI Conference on Web and Social Media (2016).
[22]S. Aral and P. S. Dhillon, "Social influence maximization under empirical influence models," Nature Human Behaviour **2**, 375 (2018).
[23]M. Kitsak, L. K. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. E. Stanley, and H. A. Makse, "Identification of influential spreaders in complex networks," Nature Physics **6**, 888 (2010).
[24]A. Banerjee, A. G. Chandrasekhar, E. Duflo, and M. O. Jackson, "The diffusion of microfinance," Science **341**, 1236498 (2013).
[25]F. Morone and H. A. Makse, "Influence maximization in complex networks through optimal percolation," Nature **524**, 65–68 (2015).
[26]L. Lü, D. Chen, X.-L. Ren, Q.-M. Zhang, Y.-C. Zhang, and T. Zhou, "Vital nodes identification in complex networks," Physics Reports **650**, 1–63 (2016).
[27]E. Muller and R. Peres, "The effect of social networks structure on innovation performance: A review and directions for research," International Journal of Research in Marketing **36**, 3–19 (2019).
[28]M. Medo, M. S. Mariani, A. Zeng, and Y.-C. Zhang, "Identification and modeling of discoverers in online social systems," Scientific Reports **6**, 34218 (2016).
[29]M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, "Understanding individual human mobility patterns," Nature **453**, 779 (2008).
[30]C. Song, Z. Qu, N. Blumm, and A.-L. Barabási, "Limits of predictability in human mobility," Science **327**, 1018–1021 (2010).
[31]N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," Science **328**, 1029–1031 (2010).
[32]M. Tribus, Thermostatics and thermodynamics: an introduction to energy, information and states of matter, with engineering applications (van Nostrand, 1961).
[33]R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, and M. C. González, "Sequences of purchases in credit card data reveal lifestyles in urban populations," Nature communications **9** (2018).
[34]D. M. Powers, "Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation," Journal of Machine Learning Technologies **2**, 37–63 (2011).
[35]E. M. Rogers, Diffusion of innovations (Simon and Schuster, 2010).
[36]S. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse, "Inferring personal economic status from social network location," Nature Communications **8**, 15227 (2017).
[37]M. J. Salganik, P. S. Dodds, and D. J. Watts, "Experimental study of inequality and unpredictability in an artificial cultural market," Science **311**, 854–856 (2006).
[38]M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in Fourth International AAAI Conference on Weblogs and Social Media (2010).
[39]D. R. Bell, S. Gallino, and A. Moreno, "Offline showrooms in omnichannel retail: Demand and operational benefits," Management Science **64**, 1629–1651 (2017).
[40]D. R. Bell, S. Gallino, and A. Moreno, "The store is dead—long live the store," MIT Sloan Management Review **59**, 59–66 (2018).
[41]S. Aral and D. Walker, "Identifying influential and susceptible members of social networks," Science **337**, 337–341 (2012).
[42]C. Catalini and C. Tucker, "When early adopters don't adopt," Science **357**, 135–136 (2017).
[43]A. M. Mood, "Introduction to the theory of statistics." (1950).
[44]H. Liao, M. S. Mariani, M. Medo, Y.-C. Zhang, and M.-Y. Zhou, "Ranking in evolving complex networks," Physics Reports **689**, 1–54 (2017).
[45]L. Pappalardo, F. Simini, S. Rinzivillo, D. Pedreschi, F. Giannotti, and A.-L. Barabási, "Returners and explorers dichotomy in human mobility," Nature communications **6** (2015).
[46]G. Vaccario, M. Medo, N. Wider, and M. S. Mariani, "Quantifying and suppressing ranking bias in a large citation network," Journal of Informetrics **11**, 766–782 (2017).
[47]E. Sarigöl, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," EPJ Data Science **3**, 9 (2014).
[48]G. James, D. Witten, T. Hastie, and R. Tibshirani, An introduction to statistical learning, Vol. 112 (Springer, 2013).
[49]D. Chicco, "Ten quick tips for machine learning in computational biology," BioData Mining **10**, 35 (2017).
[50]S. McGee, "Simplifying likelihood ratios," Journal of General Internal Medicine **17**, 647–650 (2002).

[51]S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using matthews correlation coefficient metric," PLOS ONE **12**, e0177678 (2017).
[52]B. W. Matthews, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme," Biochimica et Biophysica Acta (BBA)- Protein Structure **405**, 442–451 (1975).

Category	MCC	Description
Eating places	5812	Restaurants or eating places
	5813	Drinking Places (Alcoholic Beverages), Bars, Taverns, Cocktail lounges, Nightclubs and Discotheques
	5814	Fast Food Restaurants
Food stores	5411	Grocery stores, Supermarkets
	5422	Freezer and Locker Meat Provisioners
	5441	Candy, Confectionery, Nut stores
	5451	Dairy Products stores
	5462	Bakeries
	5499	Misc. Food stores – Convenience stores and Specialty Markets
Clothing stores	5611	Men's and Boy's Clothing and Accessories stores
	5621	Women's Ready-to-Wear stores
	5631	Women's Accessory and Specialty stores
	5641	Children's and Infant's Wear stores
	5651	Family Clothing stores
	5655	Sports Apparel, Riding Apparel stores
	5661	Shoe stores
	5681	Furriers and Fur stores
	5691	Men's and Women's Clothing stores
	5697	Tailors, Seamstress, Mending, and Alterations
	5698	Wig and Toupee stores
	5699	Miscellaneous Apparel and Accessory stores

TABLE S1. Definition of the three store categories analyzed in this paper in terms of the Merchant Category Codes (MCCs) of the included stores.

SUPPLEMENTARY NOTE S1: DATA FILTERING AND NETWORKS CONSTRUCTION

Before describing the data, we point out that all the data analyzed in the article are anonymized. The subjects of the analysis (individuals and stores) are represented by meaningless hashes in the dataset. All individuals are non-identifiable, meaning that there is no way to reconstruct the individuals' real identities; all stores are nameless, there is no way to reconstruct the stores' real name; all transactions are innominate. For the sake of better readability, when referring to the analyzed datasets and the individuals in the following, we shall omit the labels "anonymized" and "non-identifiable". Nevertheless, whenever we will refer to an individual, it should be interpreted as a non-identifiable individual whose real identity is impossible to identify. Whenever we will refer to a group of "detected" individuals, it should be interpreted as a group of individuals whose real identities cannot be identified.

Credit Card Record (CDR)

We analyzed a Credit Card Record (CCR) from a large bank in an emerging country collected over a three-year period from June 2015 to June 2018. We filtered out stores with less than ten customers throughout the whole data time span. We consider three store categories: eating places, food stores, and clothing stores. The three categories have been selected according to the Merchant Category Codes (MCCs) that are available in the data, according to the classification scheme reported in Table S1. We study separately three temporal bipartite networks where individuals are connected to the stores they purchased in. The time-stamp of each link is determined by the time-stamp of the first purchase by the individual.

We split the three-year CCR into three non-overlapping periods, as explained below. The time periods reported below refer to the ones that were used for the analysis in the main text; in Supplementary Note S5, we tested the robustness of our results with respect to other choices for the data partitioning.

- **Pre-filtering period (June 2015 – November 2015)**. We used this period to assess whether a store that appears in the training period is a new store or a previously-existing one. If a store found in the training period is also found in the pre-filtering period, it is a pre-existing one and does not contribute to the customers' number of discoveries, whereas it still contributes to their number of visited stores.

- **Training period (December 2015 – May 2017)**. We analyze separately the three categories of stores reported in Table S1. A potential issue is that the discoverer detection procedure requires us to estimate the success of the stores, which might be unreliable for stores that received their first transaction near the end of the

training period. For this reason, we filtered out from the analysis the stores introduced less than two months before the end of the training period. The total number of relevant stores is denoted as S. The time $t_{i\alpha}$ of each link (i, α) is determined by the first visit of individual i to store α. We denote as k_i and v_α the number of stores visited by individual i within the training period (excluding the last two months of this period) and the number of first-time customers of store α, respectively. We denote as d_i the number of discoveries collected by individual i within the training period (excluding the last two months of this period). The reason for excluding the last two months when measuring k_i and d_i is that the estimation of success might be unreliable for the shops that appeared for the first time near the end of the training period.

- **Validation period (June 2017 – May 2018)**. The transactions from June 2017 to May 2018 are used as validation period to assess the out-of-sample predictive signal for different groups of detected individuals. We focus on the stores that were opened within this period, and assess the relation between the presence/absence of different groups of individuals among the earliest customers and the future success of the store (see Supplementary Note S3 for the details). Again, a potential issue is that the prediction evaluation procedure requires us to estimate the future success of the stores, which might be unreliable for stores opened near to the end of the validation period. For this reason, we filtered out from the analysis the stores introduced less than two months before the end of the validation period. We denote by V the resulting number of relevant stores.

Call Data Record (CDR) and its relation with the CCR

We analyzed a Call Data Record (CDR) from a large mobile phone operator from the same country where the CCR was recorded. The CDR used in this study covers a one-year period from January 2016 to December 2016. Importantly, this period overlaps with the CCR's time span, and it is possible to partially match the individuals in the CDR with the individuals in the CCR (see below). For each telco customer, the CDR contains all the calls she made to or received from both other telco customers and non-customers. This implies that for the telco customers, we can observe their complete mobile-phone communication activity, whereas for the non-customers, we can only see their communications with the telco customers. Besides, each individual may be telco customer only for specific months of the year, but not throughout the whole year.

In our work, we use the CDR to construct 12 snapshots of the social network. For each month, we only include the telco customers in the network. When computing the time-averaged centrality metrics of each individual in the social network, we only include the months when the target individual was a telco customer. The rationale is that the individuals' centrality is largely underestimated in the months when they are not telco customers, because their calls from/to non-telco customers are not included. We refer to Supplementary Note S2 for a description of how the time-averaged centrality metrics were computed.

SUPPLEMENTARY NOTE S2: INDIVIDUAL-LEVEL TRAITS EXTRACTED FROM THE CCR

We describe here how we extracted the individual-level traits of interest from the available CCR.

- **Surprisal (Discoverers).**

  The individuals' surprisal is defined by Eq. (2) in the main text. We refer to the main text for the details of its computation. The **discoverers** of a given category of stores are the top-τ% individuals by the surprisal obtained by analyzing the respective category of stores.

- **Number of visited stores (Store explorers).**

  For each of the three categories of stores considered in our work, we count the number of visited stores per individual, k, within the training period. The **store explorers** of a given category of stores are the top-τ% individuals by number of visited stores within the training period.

- **Time-averaged total expenditure (High-expenditure individuals).**

  For each bank customer who made at least one purchase within the training period, we extract his/her total expenditures from each month between January and December 2016, and we average over this 12-month period. The **high-expenditure individuals** are the top-τ% individuals by time-averaged total expenditure.

SUPPLEMENTARY NOTE S3: INDIVIDUAL-LEVEL TRAITS EXTRACTED FROM THE CDR

We describe here how we extracted the individual-level traits of interest from the available CDR. The extracted traits are used both to detect the top-individuals used to make predictions (as detailed below), and to characterize the groups of top-individuals in Fig. 4 of the main text.

Centrality and social hubs

We introduce here appropriate time averages of three different centrality metrics: degree, collective influence, social diversity.

- **Time-averaged number of contacts (Social hubs by degree).**

  The number of contacts (or degree) of the individuals in the social network is probably the simplest metric to quantify individuals' centrality[44]. Individuals with a large number of contacts – *social hubs* – have been first used for success prediction by Goldenberg *et al.*[8]. To detect the social hubs[8], we measure the number of contacts per individual within each month, $k_i(t)$. We denote by $\mathcal{T}(t)$ the set of telco customers in month t. An individual may be telco customers only for some months within the CDR timespan. Motivated by the lines of reasoning in Supplementary Note S1, we define the time-averaged number of contacts for individual i as

$$\overline{k_i} = \frac{\sum_t \delta(i \in \mathcal{T}(t))\, k_i(t)}{\sum_t \delta(i \in \mathcal{T}(t))}. \tag{3}$$

  From the definition, it follows that $\overline{k_i} > 0$ only for those individuals who are telco customers for at least one month. Only the months when an individual is a telco customer are included in the average. The *social hubs by degree* are the top-τ% individuals by $\overline{k_i}$, among the individuals who are found in the CCR and made at least one purchase in stores of the analyzed category within the training period.

  We note that if an individual is found in the CCR but she is not among the telco customers, she obtains $\overline{k}_i = 0$ and cannot be detected as a social hub. This is a consequence of the incompleteness of our data: given a set of individuals who make transactions in the CDR, we do not know the communication activity for all of them and, as a result, we can only detect the top-individuals by centrality among those that are also telco customers. Similar remarks apply for all other individuals' traits extracted from the CDR.

- **Time-averaged Collective Influence (Social hubs by Collective Influence, CI).**

  While the degree is the simplest centrality metric in networks[44], it neglects higher-order network effects that are potentially informative about the position of the nodes. As a more sophisticated metric of network centrality, we rely on the *collective influence* metric introduced by Morone and Makse[25]. The metric detects the minimal set of nodes that, once removed from the network, disrupt the network's giant component. The detection of these nodes is typically referred to as the structural influence maximization problem[26]. Morone and Makse[25] solved analytically the problem through the theory of optimal percolation on graphs, and showed that the collective influence metric provides a reliable approximation to the solution of the problem and, at the same time, can be computed rapidly on large datasets. By considering the network of telco customers in month t, the collective influence $CI_i(t)$ of a telco customer i in month t is given by[25]

$$CI_i^{(l)}(t) = (k_i(t) - 1) \sum_{j \in \partial B_l(i)} (k_j(t) - 1), \tag{4}$$

  where $B_l(i)$ denotes the ball of radius l centered in i, and $\partial B_l(i)$ denotes its frontier[25]. Here, we set $l = 1$. As we did for the number of contacts, we define the time-averaged collective influence of individual i as

$$\overline{CI_i^{(1)}} = \frac{\sum_t \delta(i \in \mathcal{T}(t))\, CI_i^{(1)}(t)}{\sum_t \delta(i \in \mathcal{T}(t))}. \tag{5}$$

  The *social hubs by collective influence* are the top-τ% individuals by $\overline{CI_i^{(1)}}$, among the individuals who are also found in the CCR and made at least one purchase in stores of the analyzed category within the training period.

- **Time-averaged social diversity (Social hubs by social diversity)**

We consider an alternative metric of social importance that has brought insights into regional socioeconomic development[31]. The social diversity metric quantifies whether a given individual tends to communicate repeatedly with a restricted set of contacts, or whether she contacts a diverse set of people. It is defined as the entropy[31]

$$S_i^S(t) = -\sum_j \frac{(w_{ij}(t)/w_i(t))\,\log(w_{ij}(t)/w_i(t))}{\log(w_i(t))} \tag{6}$$

where w_i represents i's total number of interactions within month t, $w_{ij}(t)$ the total number of interactions between i and j within month t. If a person i has only one contact j over one month t, then $w_{ij}(t) = \delta_{ij}$ and, as a consequence, $S_i^S(t) = 0$. As we did for the previous centralities, we define the time-averaged social diversity of individual i as

$$\overline{S_i^S} = \frac{\sum_t \delta(i \in \mathcal{T}(t))\, S_i^S(t)}{\sum_t \delta(i \in \mathcal{T}(t))}. \tag{7}$$

The *social hubs by social diversity* are the top-τ% individuals by $\overline{S^S}$, among the individuals who are also found in the CCR and made at least one purchase in stores of the analyzed category within the training period.

Mobility-related traits and explorers

- **Time-averaged mobility diversity**

The mobility diversity metric has been introduced to characterize the predictability of human mobility[30]. The idea is to understand whether a given individual uses a restricted number of antennas or a diverse set of antennas. The latter means that the individual has been in a diverse set of locations, which is a manifestation of high mobility. The mobility diversity of individual i is given by the following entropy[30]:

$$S_i^M(t) = -\sum \frac{(c_{ia}(t)/c_i(t))\,\log(c_{ia}(t)/c_i(t))}{\log(c_i(t))} \tag{8}$$

where $c_{ia}(t)$ denotes the number of calls made/received by i through antenna a within month t, and c_i denotes the total number of calls made/received by individual i within month t. If i uses only one antenna within month t, $c_{ia}(t) = c_i(t)$ which results in $S_i^M(t) = 0$. On the other hand, individuals who made/received calls from many different antenna are characterized by large values of $S_i^M(t)$. As we did for the previous CDR-extracted traits, we define the time-averaged mobility diversity of individual i as

$$\overline{S_i^M} = \frac{\sum_t \delta(i \in \mathcal{T}(t))\, S_i^M(t)}{\sum_t \delta(i \in \mathcal{T}(t))}. \tag{9}$$

The *explorers by mobility diversity* are the top-τ% individuals by $\overline{S^M}$, among the individuals who are also found in the CCR and made at least one purchase in stores of the analyzed category within the training period.

- **Time-averaged radius of gyration**

The radius of gyration can be interpreted as the characteristic distance traveled by a given individual[29]. This metric has been used to distinguish explorative individuals from "returner" individuals who tend to only visit a small number of locations[45]. Individual i's total radius of gyration r_g is defined as[29]

$$r_i = \sqrt{\frac{1}{N_i}\sum_{l\in\mathcal{L}_i} n_{il}\left(\mathbf{r}(l) - \mathbf{r}_i^{CM}\right)^2}, \tag{10}$$

where n_{il} denotes the number of times individual i uses antenna l within month t, $\mathcal{L}_i$ denotes the set of antennas that individual i visited, $N_i = \sum_{l\in\mathcal{L}} n_i$, $\mathbf{r}(l)$ is a two-dimensional vector describing the geographic coordinates of location l, $\mathbf{r}_i^{CM}$ represents individual i's center of mass. As we did for the previous CDR-extracted traits, we define the time-averaged radius of gyration of individual i as

$$\overline{r_i} = \frac{\sum_t \delta(i \in \mathcal{T}(t))\, r_i(t)}{\sum_t \delta(i \in \mathcal{T}(t))}. \tag{11}$$

The *explorers by radius of gyration* are the top-τ% individuals by $\overline{r}$, among the individuals who are also found in the CCR and made at least one purchase in stores of the analyzed category within the training period.

SUPPLEMENTARY NOTE S4: SUCCESS PREDICTION

We provide here all details on the formulation of the predictive problem, the Naive Bayes Classifiers used in the paper, and the prediction evaluation metrics.

Formulation of the predictive problem

In line with the literature on the prediction of popularity in online systems[10,11,21], we aim to address the following question: Can we use the behavioral and socioeconomic traits of stores' early customers to predict the eventual popularity of the store? In other words, we *peek* into the stores' early activity data, and we aim to use this information to predict the stores' eventual success. Formulating the related predictive problem requires two design choices: (1) How much peeking into early activity in a store is allowed, and (2) which metric of store success we aim to predict[11,21].

As for (1), ideally, we would like to look at a period of early activity that is short enough that the eventual success of the store is not evident. Besides, we would like to exclude from the predictive problem stores that only received few customers and, for this reason, are unlikely to become successful in the future. Motivated by these considerations, we peek into the stores' earliest w first-time customer, where we vary w from 5 to 30; when we consider the earliest w first-time customer, only stores that received at least w first-time customer are included in the analysis.

As for (2), a broad range of predictive problems emerges depending on the dimension of store success that we aim to predict. In this work, we focus on the stores' popularity, v, defined by their total number of first-time customers (i.e., the total number of people who purchase in the store at least once). However, we cannot use directly v to define the group of successful stores, as v is strongly influenced by store age and MCC (see Supplementary Note S2). One faces an analogous issue when measuring the impact of scientific papers through citation-based indicators (which often results in metrics of impact that are biased by paper age and scientific field[44,46]), or when attempting to compare the performance of different innovation diffusion processes that started at different points in time[27]. To factor out these confounding effects, we define the group $\mathcal{P}$ of popular stores as the group of stores that are ranked among the top-s% stores by cumulative number of first-time customers, *among stores with the same MCC that received their first transaction in the same month.* The cumulative number of first-time customers is measured throughout the complete validation period. Assessing the popularity of each store only in relation to stores of the same MCC and age directly removes the bias of v without the need to know the distribution of v.

Putting together (1) and (2), we formulate a binary classification problem where we observe the earliest w first-time customers of a store within the validation period (excluding the last two months), and we aim to predict whether the store will end up in the group $\mathcal{P}$ of the popular stores. Importantly, to validate the predictions, we only consider stores that received their first purchase during the validation period (excluding the last two months), whereas the eight different classes of top-individuals defined in the main text and Supplementary Note S2 are detected within the training period.

Naive Bayes Classifiers (NBCs)

To quantify the predictive performance of different groups of individuals, we consider **Naive Bayes Classifiers** (NBCs)[47]. For each group $\mathcal{I}$ of relevant individuals (e.g., $\mathcal{I}$ can represent the set of discoverers), we consider the simplest possible classification rule: *a store is classified as successful if at least one individual $i \in \mathcal{I}$ was among the earliest w customers, as non-successful otherwise.* Such a simple classifier allows us to compare the predictive power of the eight groups of relevant individuals considered here. From a machine-learning standpoint, this classifier requires no model training: it only takes as input the list of individuals detected within the training period. It is useful to introduce a random classifier that provides us, for each prediction evaluation metric, with a baseline performance metric. Such a random classifier is simply a random guess: each store is classified as successful with probability 0.5, as non-successful with probability 0.5.

Beyond this simple classifier, we also consider "two-dimensional" NBCs that result from the joint presence of individuals from two groups of relevant individuals. For each pair $(\mathcal{I}_1,\mathcal{I}_2)$ of relevant individuals (e.g., $\mathcal{I}_1$ and $\mathcal{I}_2$ may represent the set of discoverers and social hubs, respectively), we implement the following classification rule: *a store is classified as successful if and only if at least one individual from $\mathcal{I}_1$ and one individual from $\mathcal{I}_2$ were among the earliest V first-time customer, as non-successful otherwise.* Compared to the 1-dimensional NBCs, this rule is stricter, which results in a smaller number of stores classified as successful. As shown in Fig. 3 of the main text, this can result in precision improvements that range from marginal to large, depending on the pairs of considered individuals and store category.

Prediction evaluation metrics

The performance in the classification task is measured in terms of the four elements of the confusion matrix (also referred to as contingency table)[34,48,49]: number of True Positives (T^+), False Positives (F^+), True Negatives (T^-), False Negatives (F^-). We consider the following evaluation metrics:

- **Precision** or **success rate**[34]. Fraction of positive-classified stores that ended up in $\mathcal{P}$. In formulas, the precision P is given by[34]

$$P = \frac{T^+}{T^+ + F^+}. \tag{12}$$

  Therefore, the ratio between the precision for the NBC of group $\mathcal{I}$ and the expected precision of the random classifier (or equivalently, the ratio between $\mathcal{I}$'s success rate and the baseline success rate) can be interpreted as the *fold increase in success rate*[4] (see main text). Groups of individuals with success-rate fold increase substantially larger than one can be interpreted as predictors of success.

- **Recall.** The recall, R, is commonly used in information retrieval and it is typically considered as a complementary metric to the precision. It is defined as the fraction of successful items that are labeled as positive[34]

$$R = \frac{T^+}{T^+ + F^-}. \tag{13}$$

  In our study, the recall metric naturally favors groups of individuals who purchased in many different stores, because they are more likely to label a store as positive and, therefore, to label the successful ones as positive. However, the metric is still informative to have a full understanding of the predictions by the different groups of individuals: While in the main text we assessed whether stores that received an early purchase by a discoverer are more likely to be successful, the recall metric informs us about how many successful stores we can hope to detect by tracking the discoverers.

- **Positive likelihood ratio.** The positive likelihood ratio, L, is commonly used in medicine and diagnostic testing[50]. For our problem, it is defined as the probability that a positive-classified store does end up in $\mathcal{P}$ divided by the probability that a negative-classified store does end up in $\mathcal{P}$. It can be expressed in terms of recall and specificity[34]:

$$L = \frac{R}{1 - S}. \tag{14}$$

  where $R = T^+/(T^+ + F^-)$ denotes the recall (or true positive rate) and $S = T^-/(T^- + F^+)$ denotes the specificity (or true negative rate).

- **Matthews' correlation coefficient (r).** As the number of popular stores, $|\mathcal{P}|$, is substantially smaller than the total number of stores, traditional prediction-evaluation metrics like accuracy and the F1-score become ineffective to evaluate classifiers' predictive performance. It has been shown numerically[51?] that the Matthews' correlation coefficient is substantially less sensitive to data imbalance than the accuracy and the F1-score and, therefore, it is preferable[49]. The Matthews' correlation coefficient, r, is defined by the equation[52]

$$r = \frac{T^+ T^- - F^+ F^-}{\sqrt{(T^+ + F^+)\,(T^+ + F^-)\,(T^- + F^+)\,(T^- + F^-)}}. \tag{15}$$

  Differently from the accuracy and F1 score, r is robust with respect to variations of relative class size[51], which makes it suitable for our classification task that features imbalanced classes[49].

---

[4] The terminology "fold change" is typically used in biology. In our study, it is particularly convenient because it describes effectively the fact that the stores with individuals from a given group of individuals among their early customers may exhibit increased odds of success.

Results

The full results for the parameters adopted in the main text are reported in Fig. S6, and the results for different parameter settings are reported in Figs. S7–S13 and discussed in Supplementary Note S6. Here, we briefly discuss the results in Fig. S5. The results for the success-rate fold increase were already shown in Fig. 2 in the main text and discussed there. The positive likelihood ratio is in almost perfect agreement with the success-rate fold increase. The Matthews' correlation coefficient is also in good agreement with the success-rate fold increase, although some evident deviations can be observed: for example, store explorers and high-expenditure individuals emerge clearly as the second best-performing group of individuals for eating places and food stores, respectively. The results for recall significantly differ, which reflects the fact that metric rewards highly-active individuals who purchase in many different stores. The store explorers emerge as the best-performing group, followed by the discoverers and the high-expenditure individuals.

In the main text, we focused on the simplest possible predictive classifier based on the early purchases by different groups of individuals. We report here that precision improvements are possible through combinations of these features. The two-dimensional classifiers introduced above tend indeed to outperform the one-dimensional classifier based on the discoverers (Fig. S7). For example, combining the discoverers with one of the three groups of explorers or high-expenditure individuals leads to a success-rate improvement for all three store categories. On the other hand, combining the discoverers with one of the three groups of social hubs leads to a success-rate improvement only for eating places and food stores, but not for clothing stores. This is unsurprising given the poor performance of the three groups of social hubs when considered independently (Fig. 2 in main text), and it suggests that combining pairs of groups of individuals can marginally improve the success rate, but only if both groups have an above-baseline success rate when considered alone.

SUPPLEMENTARY NOTE S5: ROBUSTNESS OF THE PREDICTIONS

Our predictive analysis has five main parameters. Two parameters are specific to the discoverers detection method (Δ, z), and three parameters are related to the predictive problem design $(s, \tau, \Delta t_{id})$. We provide below a description of the parameter variations that we performed, and an overview of the results. We begin by the two parameters of the discoverers detection method:

- The duration of the time window over which discoveries can be collected (time window for discoveries), Δ. By definition, a discovery is indeed only achieved when an individual purchases in a store no more than Δ days after the store received its first transaction. In the main text, we set $\Delta = 90\,dd$. We subsequently tested $\Delta = 60, 120, \infty$ (see Fig. S7 for the results). Note that with $\Delta = \infty$, a successful store can be discovered throughout the whole training period, regardless of the time at which it receives its first transaction. The results (Fig. S8) indicate that for $\Delta > 90\,dd$, the predictive power is little sensitive to Δ. On the other hand, a too narrow time window for discoveries ($\Delta = 60\,dd$) leads to suboptimal performance.

- The percentage of stores that are considered as successful in the discoverer detection procedure, $z\%$. In the main text, we set $z\% = 10\%$. We subsequently tested $z\% = 5, 20, 30, 100\%$ (see Fig. S8 for the results). Interestingly, the discoverers detected with $z\% = 100\%$ cannot be interpreted as discoverers of successful stores because with this value of $z\%$, all stores are candidates for discoveries. Therefore, the detected individuals are better interpreted as persistent early adopters. The results are mixed: For eating and food stores, a more restrictive definition of success in the training period ($z\% = 5\%$) can lead to a better out-of-sample performance in terms of precision, Matthews' correlation, and positive likelihood ratio. On the other hand, for clothing stores, a looser filter for successful stores ($z = 20\%$) and even no filter at all ($z = 100\%$) can achieve a better performance than that by the discoverers used in the main text ($z = 10\%$). This is in qualitative agreement with the finding that explorative individuals perform well for clothing stores (see Fig. 2C).

We turn our attention to the three parameters related to the predictive problem design:

- The percentage of stores that are considered as successful in the prediction evaluation, $s\%$. In the main text, we set $s\% = 10\%$. We subsequently tested $s\% = 5\%, 20\%$ (see Figs. S9–S10 for the results). We make a conservative choice and always use the same parameters for the discoverer detection as in the main text ($z\% = 10\%$); in principle, better performance might be achieved by tuning the value of $z\%$ to match $s\%$. There is still a positive predictive signal for the discoverers across the three categories. We note that the discoverers' signal weakens the most for clothing stores for a more selective success threshold ($s\% = 5\%$, Fig. S10); in this setting, the explorers by radius of gyration emerge as better predictors of success.

- The percentage of individuals selected as top-individuals by each metric, $\tau\%$. In the main text, we set $\tau\% = 1\%$. We subsequently tested a more selective threshold, $\tau\% = 0.5\%$ (see Fig. S11 for the results). The results are mostly consistent with those obtained in the main text.

- The relative duration of the training and validation period. By denoting with Δt_{id} and Δt_{val} the duration of the training and validation period, respectively, we started by performing all the analysis with $\Delta t_{id} = 18\,mm$ (from Dec. 2015 to May 2017) and $\Delta t_{val} = 12\,mm$ (from June 2017 to May 2018), as described in Supplementary Note S1. Subsequently, we repeated the analysis for a shorter training period ($\Delta t_{id} = 16\,mm, \Delta t_{val} = 14\,mm$, see Fig. S13) and a longer one ($\Delta t_{id} = 20\,mm, \Delta t_{val} = 10\,mm$, see Fig. S14). The results indicate that a longer validation period (and, correspondingly, a shorter training period) can lead to a stronger predictive power for the discoverers. In particular, in Fig. S13 and for less than 15 early customers included, the discoverers emerge as the top-performing individuals also for clothing stores. The discoverers' predictive power becomes weaker for a shorter validation period (Fig. S14), yet the discoverers remain the top-performing individuals for eating places and food stores.

To summarize, in our analysis, we started by performing the complete analysis with a predefined set of parameters ($\Delta = 90\,dd, z\% = 10\%, s\% = 10\%, \tau\% = 1\%, \Delta t_{id} = 18\,mm$, results reported in Figs. 2-3 of the main text and Fig. S6), and subsequently assessed the robustness of our predictions against variations of the parameters (Figs. S8–S14). The results are robust with respect to reasonable variations of the parameters, yet some observed variations are highly informative of the role of the various parameters. In particular, compared to the predictive results reported in the main text, factors that can improve the discoverers' predictive power are: calibrating the definition of success used in the discoverer detection (i.e., $z\%$); a longer validation period (and therefore, a larger number of stores used to validate the predictions); a more restrictive threshold ($\tau\%$) to select the top-individuals that are tracked to make the predictions.

SUPPLEMENTARY FIGURES

Basic data properties

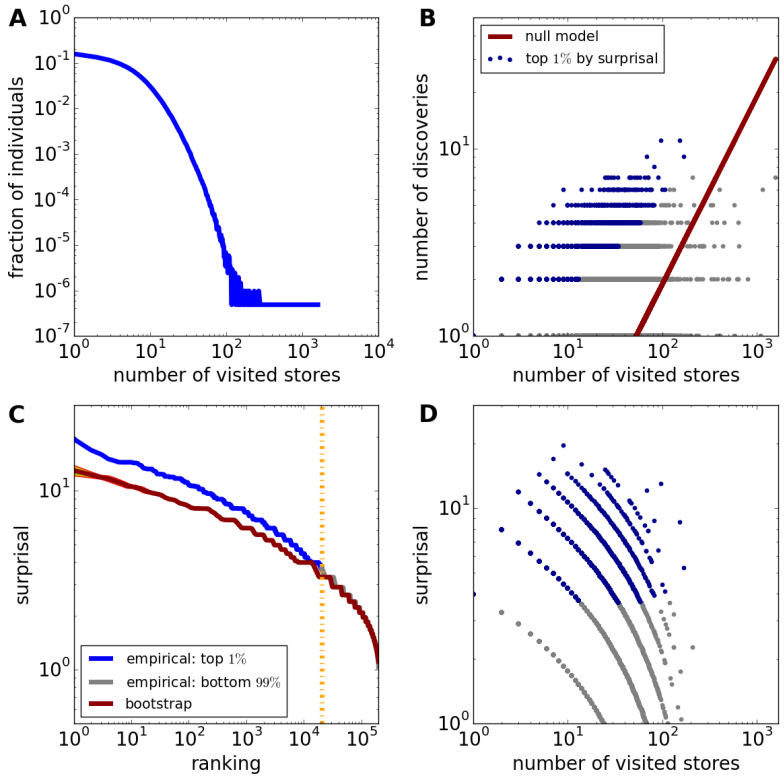

FIG. S1. **Individuals' activity and discovery patterns for food stores**. (A) Distribution of the number of visited stores per individual. (B) The number of discoveries and number of visited stores per individual are positively correlated, yet the trend deviates from the expected one under the null model, and some individuals collect significantly more discoveries than expected from their activity alone. (C) Zipf plot of the individuals' surprisal for the empirical data and the null model. The discrepancy between the two curves is significant, meaning that the surprisal values of the top-1% individuals by empirical surprisal (i.e., those at the left of the vertical dashed line) are significantly above the surprisal values expected for their ranking position. (D) The individuals' surprisal is only weakly correlated with the number of visited stores.

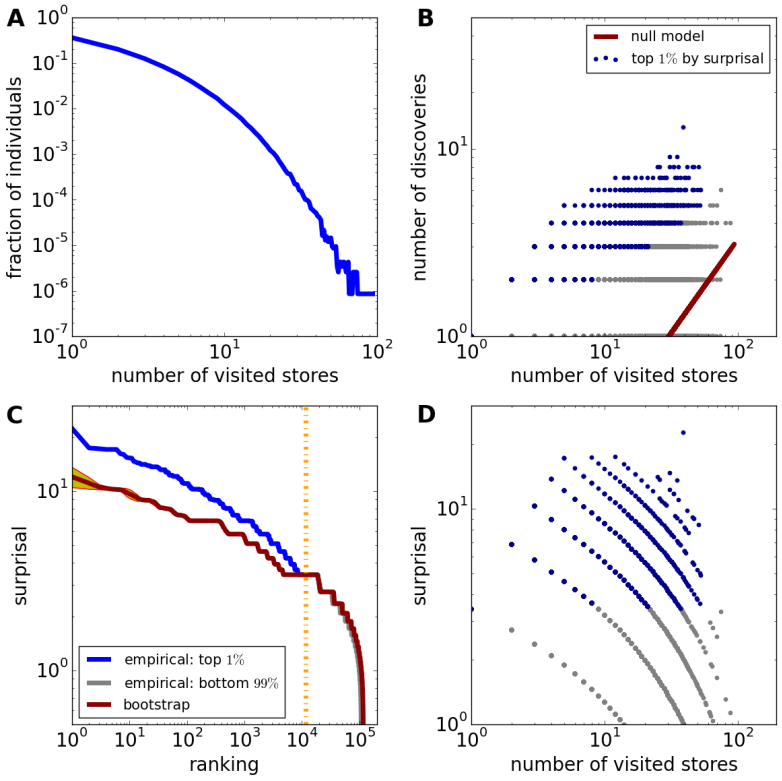

FIG. S2. **Individuals' activity and discovery patterns for clothing stores**. (A) Distribution of the number of visited stores per individual. (B) The number of discoveries and number of visited stores per individual are positively correlated, yet the trend deviates from the expected one under the null model, and some individuals collect significantly more discoveries than expected from their activity alone. (C) Zipf plot of the individuals' surprisal for the empirical data and the null model. The discrepancy between the two curves is significant, meaning that the surprisal values of the top-1% individuals by empirical surprisal (i.e., those at the left of the vertical dashed line) are significantly above the surprisal values expected for their ranking position. (D) The individuals' surprisal is only weakly correlated with the number of visited stores.

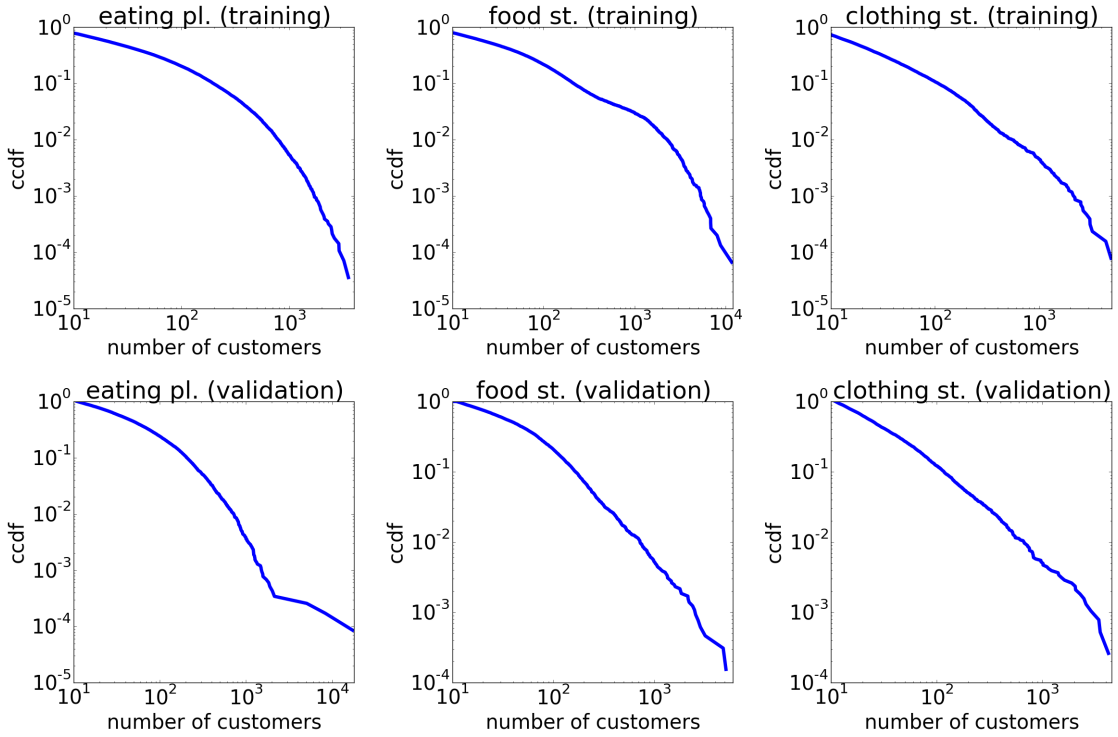

FIG. S3. **Store popularity distributions.** Complementary Cumulative Distribution Function (CCDF) of the number of first-time customers per store, c, for stores that received their first transaction within the training period (top panels) and the validation period (bottom panels), respectively. As we filtered out stores with less than 10 customers over the whole dataset, we only show the distribution in the range $c \geq 10$.

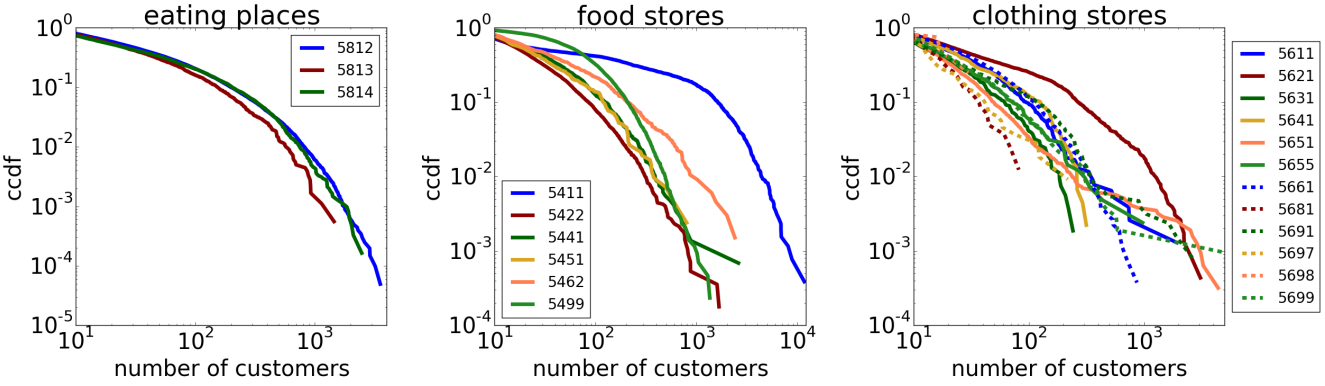

FIG. S4. **Store popularity distributions for different Merchant Category Codes (MCCs).** Complementary Cumulative Distribution Function (CCDF) of the number of first-time customers per store, for stores that received their first transaction within the training period. We plot separately the distributions for stores that belong to different Merchant Category Codes (see Table S1). The results indicate that different MCCs exhibit different success distributions, which motivates our choice to compare only stores with the same MCC to detect the group of popular stores (see Material and Methods in the main text).

FIG. S5. **Number of new stores per month.** A store is counted for a given month if it receives its first transaction on that month.

Prediction results: additional metrics

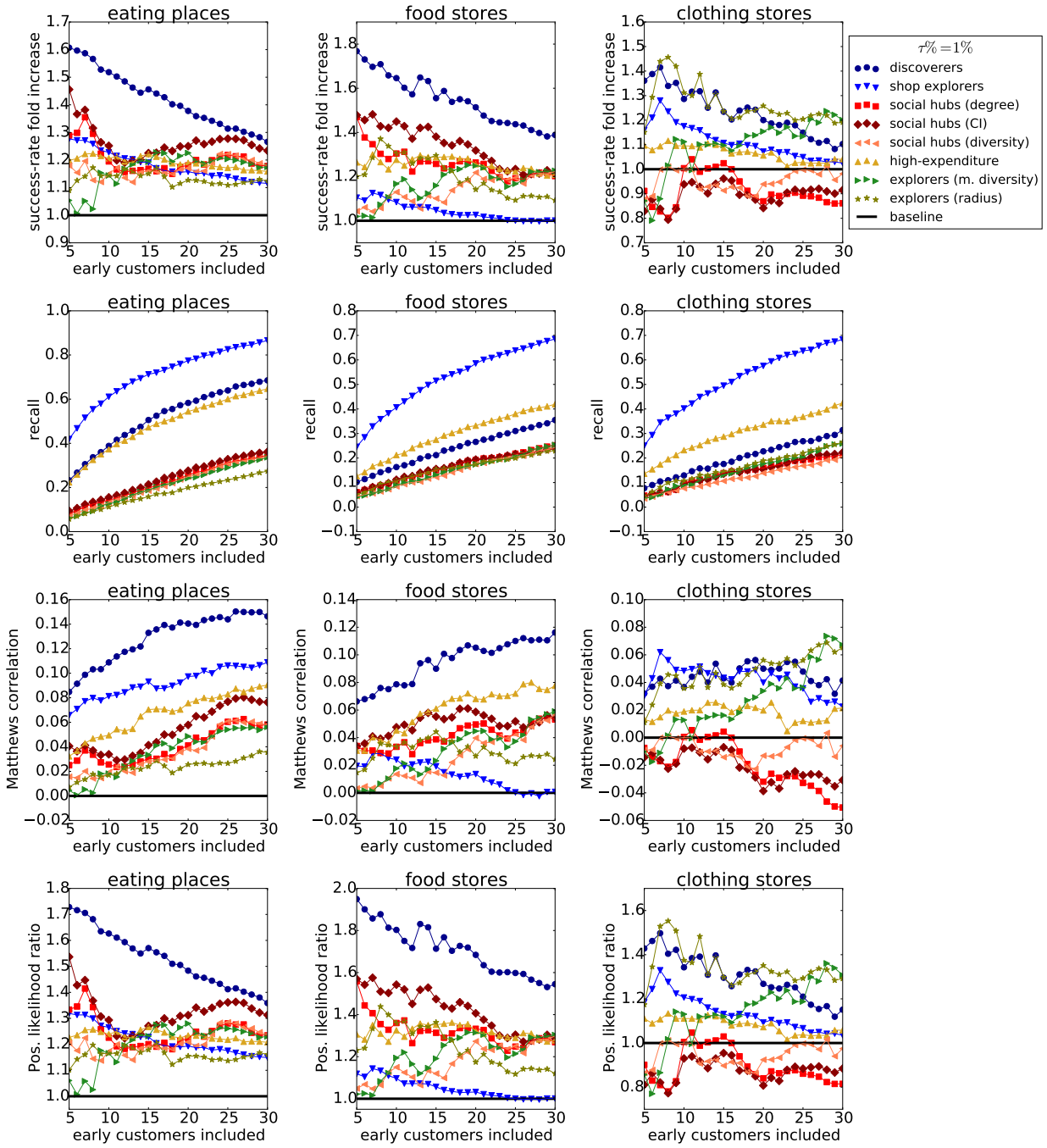

FIG. S6. **Prediction results with the parameters adopted in the main text:** $\Delta = 90, z\% = 10\%, s\% = 10\%, \tau\% = 1\%$.

Prediction results: Two-dimensional classifiers

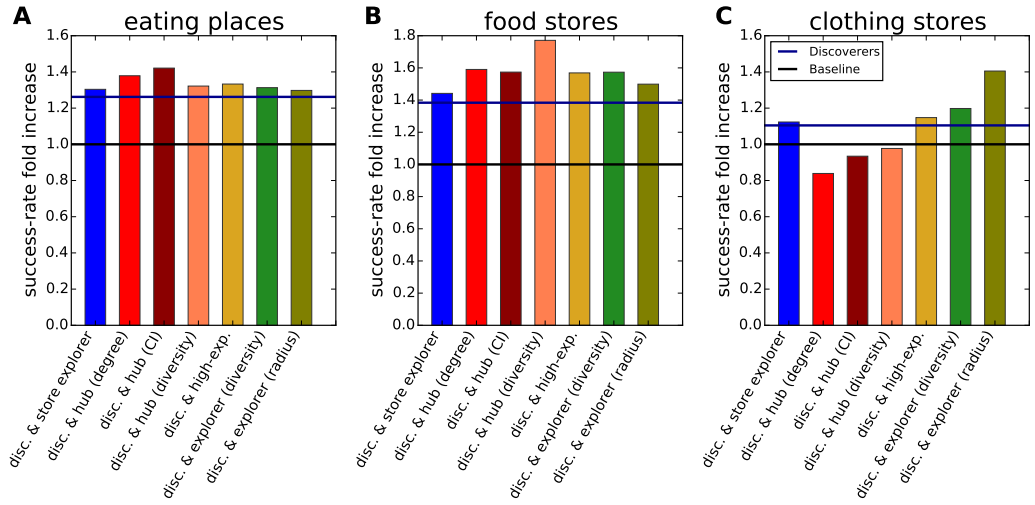

FIG. S7. **Success predictions based on pairs of groups of key individuals.** Success-rate fold increase for stores that received an early purchase by individuals that belong to two different groups of top-individuals. We include the 30 earliest customers here. Stores that received an early purchase by both a discoverer and an individual from another group of top-individuals can exhibit a substantially larger success rate than stores that received an only purchase by a discoverer. For example, stores that received an early purchase by both a discoverer and a high-expenditures individual exhibit a success-rate fold increase of $1.34, 1.58, 1.13$ for eating places, food stores, and clothing stores, respectively, representing improvements by $5.6\%, 14.9\%, 0.6\%$ with respect to the respective discoverers' success rates.

Prediction results: varying Δ

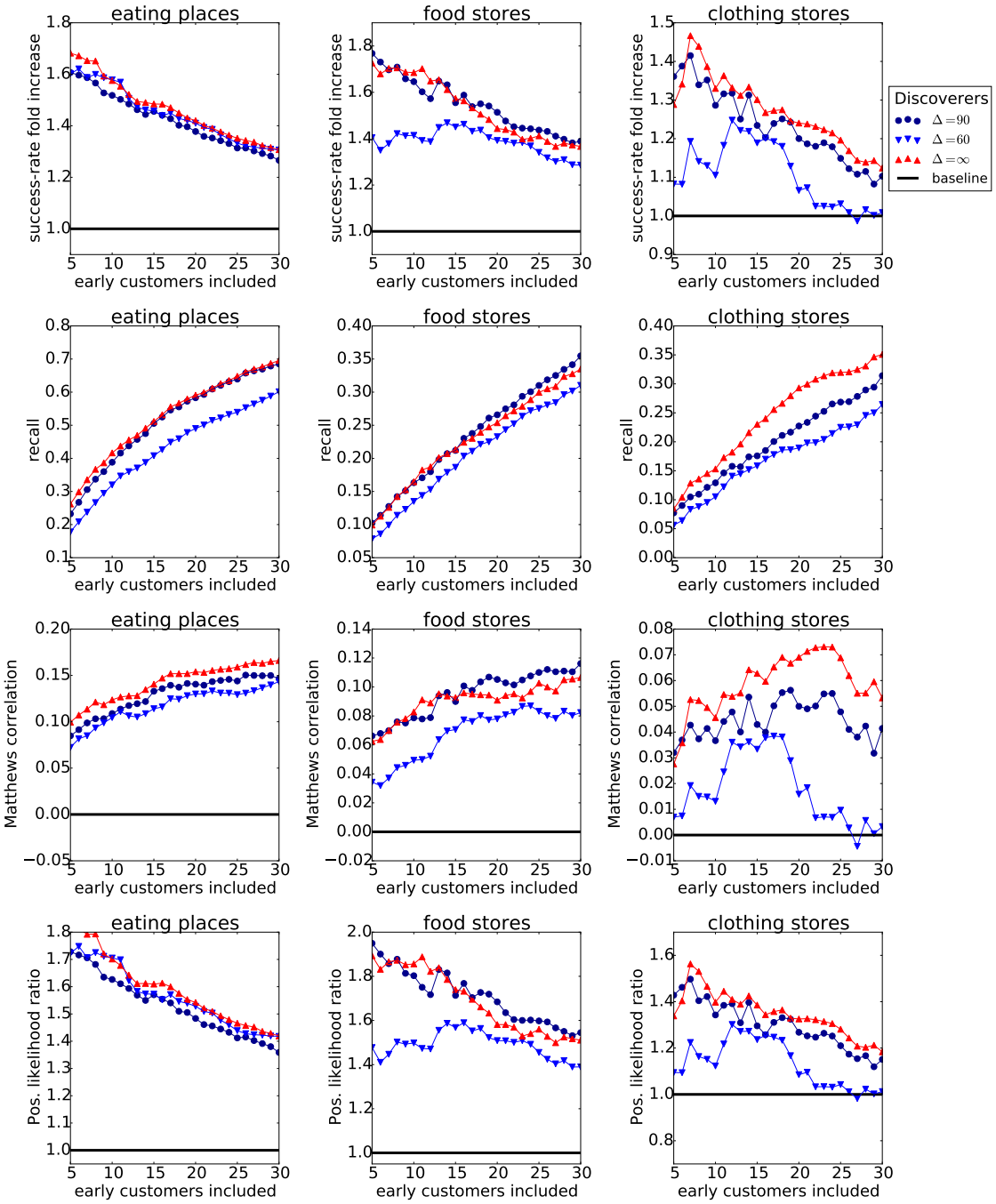

FIG. S8. **Prediction results for different values of Δ:** $\Delta = 60, 90, \infty$ (with the other parameters being fixed as in the main text: $z\% = 10\%, s\% = 10\%, \tau\% = 1\%$).

Prediction results: varying $z\%$

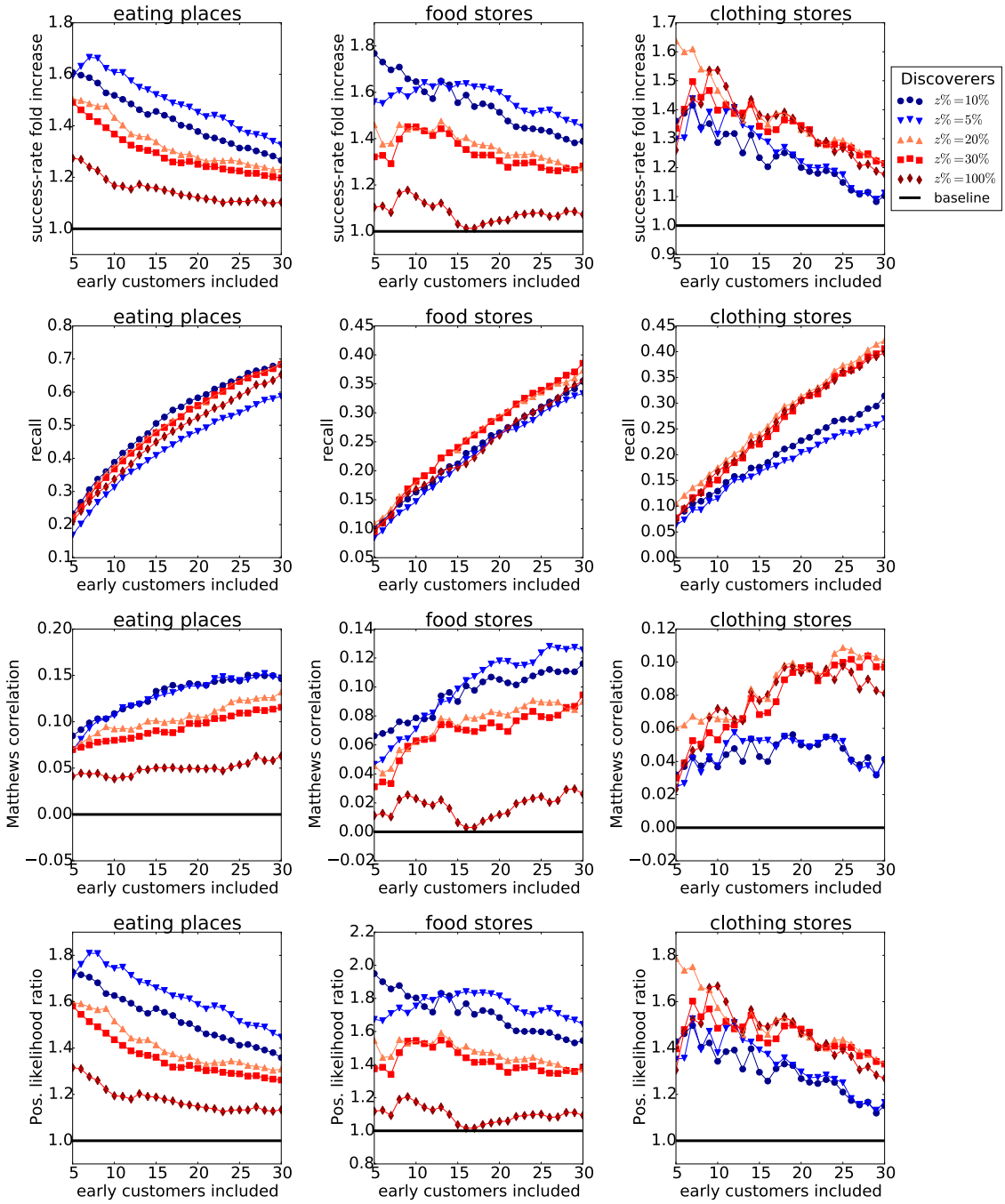

FIG. S9. **Prediction results for different values of $z\%$: $z\% = 10\%$ (main-text value), and $z\% = 5, 20, 30, 100\%$** (with the other parameters being fixed as in the main text: $\Delta = 90\,\mathrm{dd}, s\% = 10\%, \tau\% = 1\%$).

Prediction results: varying $s\%$

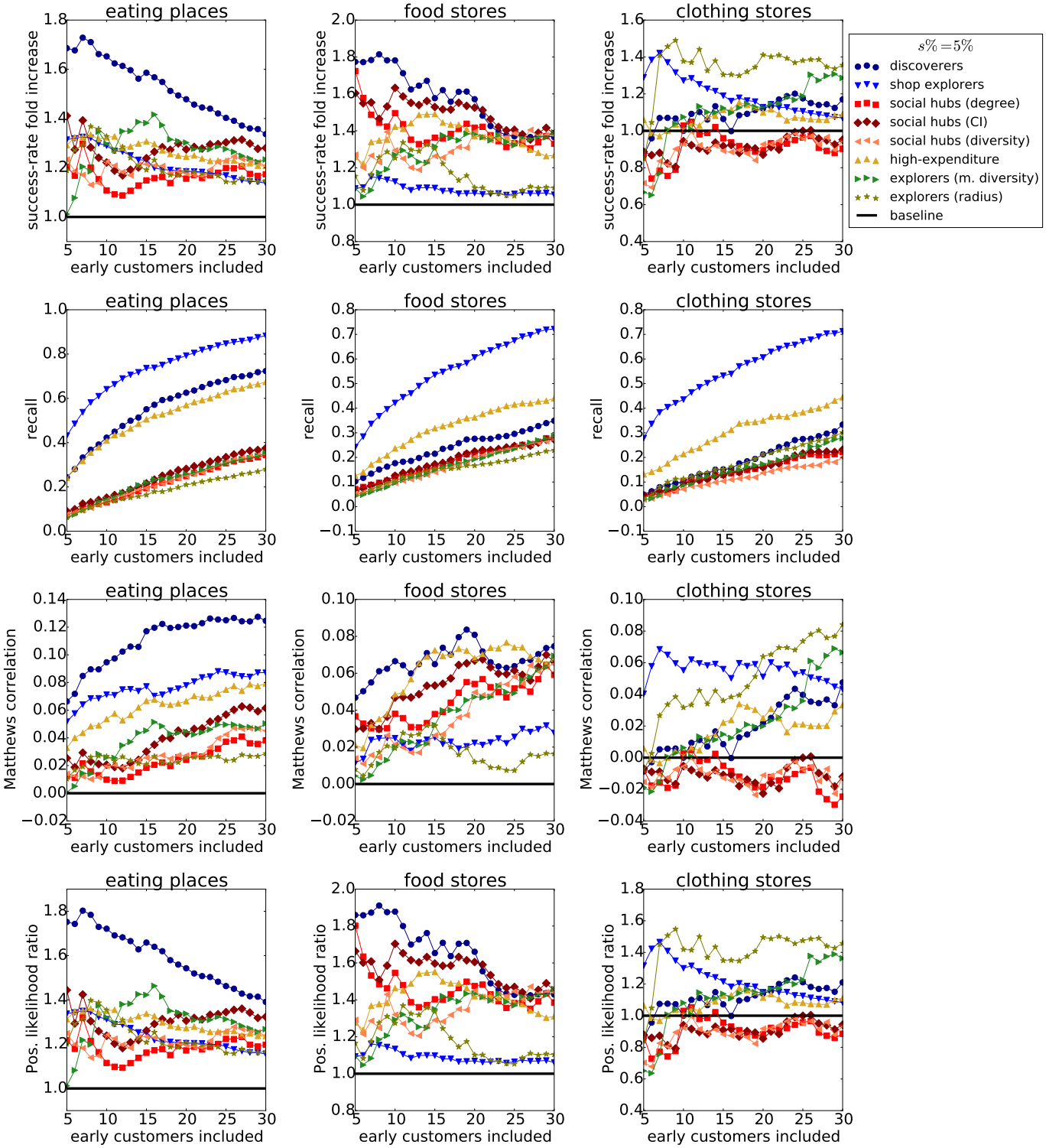

FIG. S10. **Prediction results for a more selective value of $s\%$:** $\Delta = 90, z\% = 10\%, s\% = 5\%, r\% = 0.5\%$.

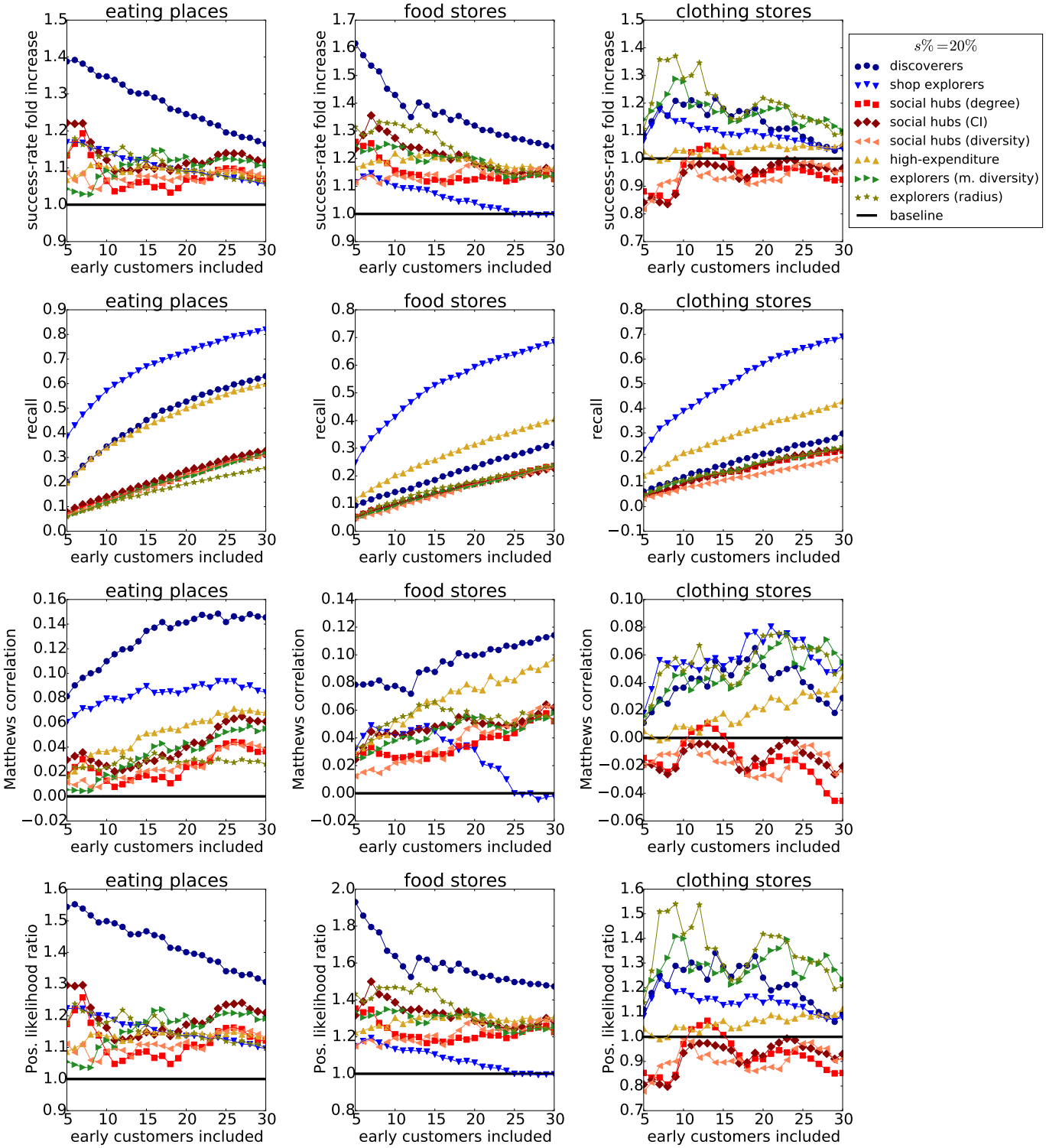

FIG. S11. **Prediction results for a less selective value of $s\%$:** $\Delta = 90, z\% = 10\%, s\% = 20\%, \tau\% = 0.5\%$.

Prediction results: varying $\tau\%$

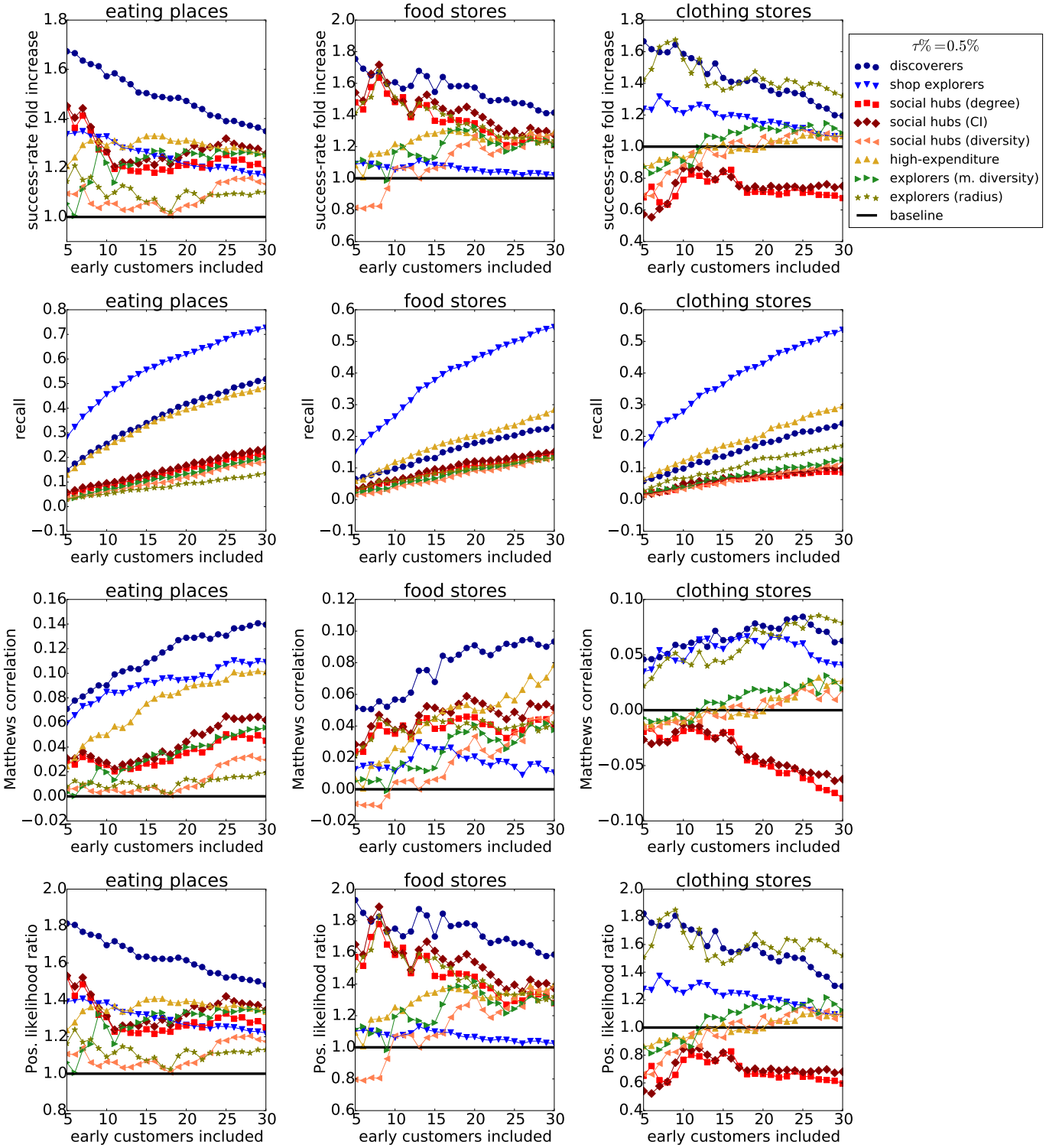

FIG. S12. **Prediction results for a more selective value of $\tau\%$:** $\Delta = 90, z\% = 10\%, s\% = 10\%, \tau\% = 0.5\%$.

Prediction results: varying the relative duration of the training and validation period

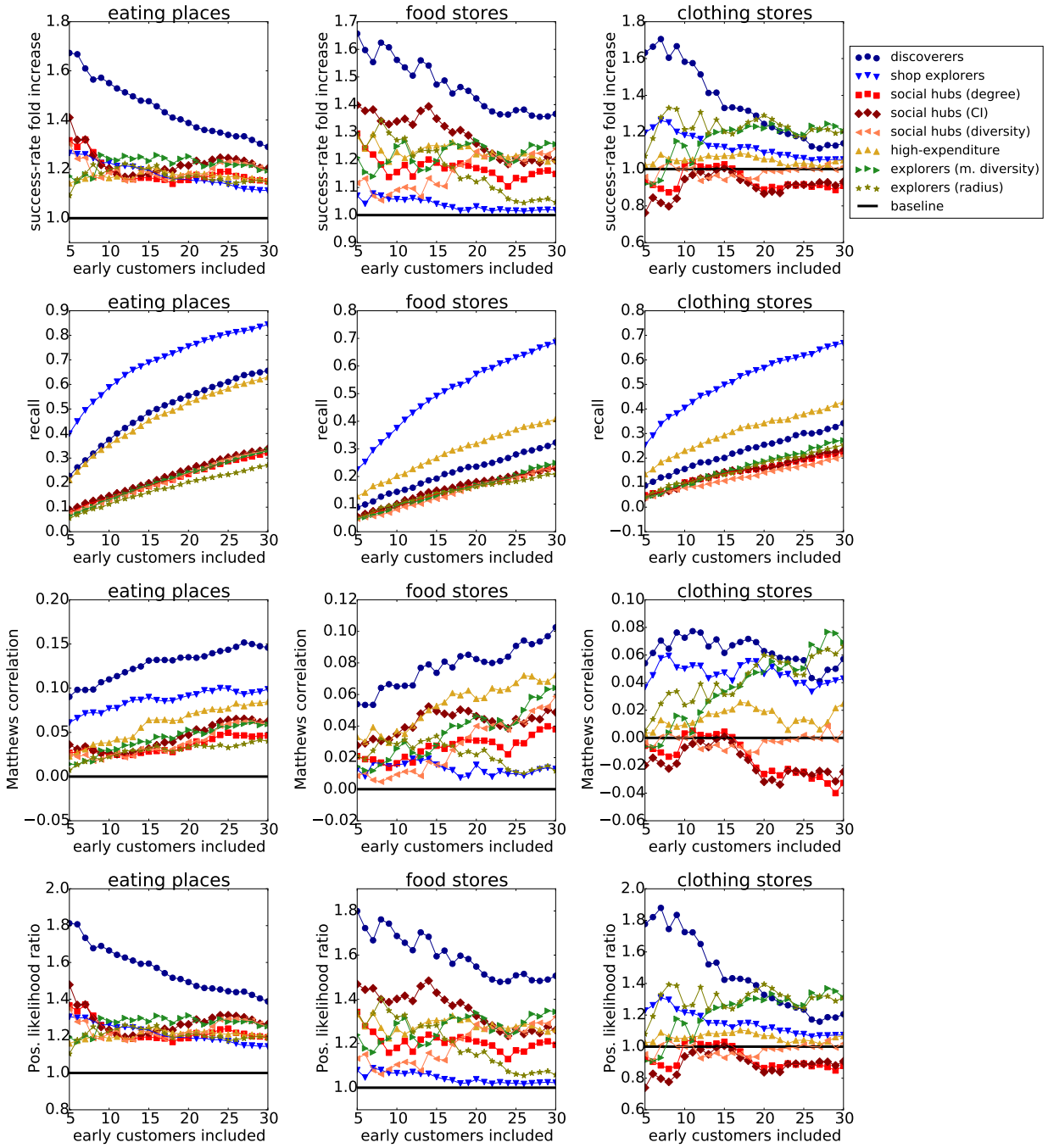

FIG. S13. **Prediction results for a shorter training period (from Dec. 2015 to Mar. 2017, 16 months) and a longer validation period (from Apr. 2017 to May 2018, 14 months) compared to those used in the main text**. The other parameters of the analysis are the same as in the main text: $\Delta = 90, z\% = 10\%, s\% = 10\%, \tau\% = 1\%$.

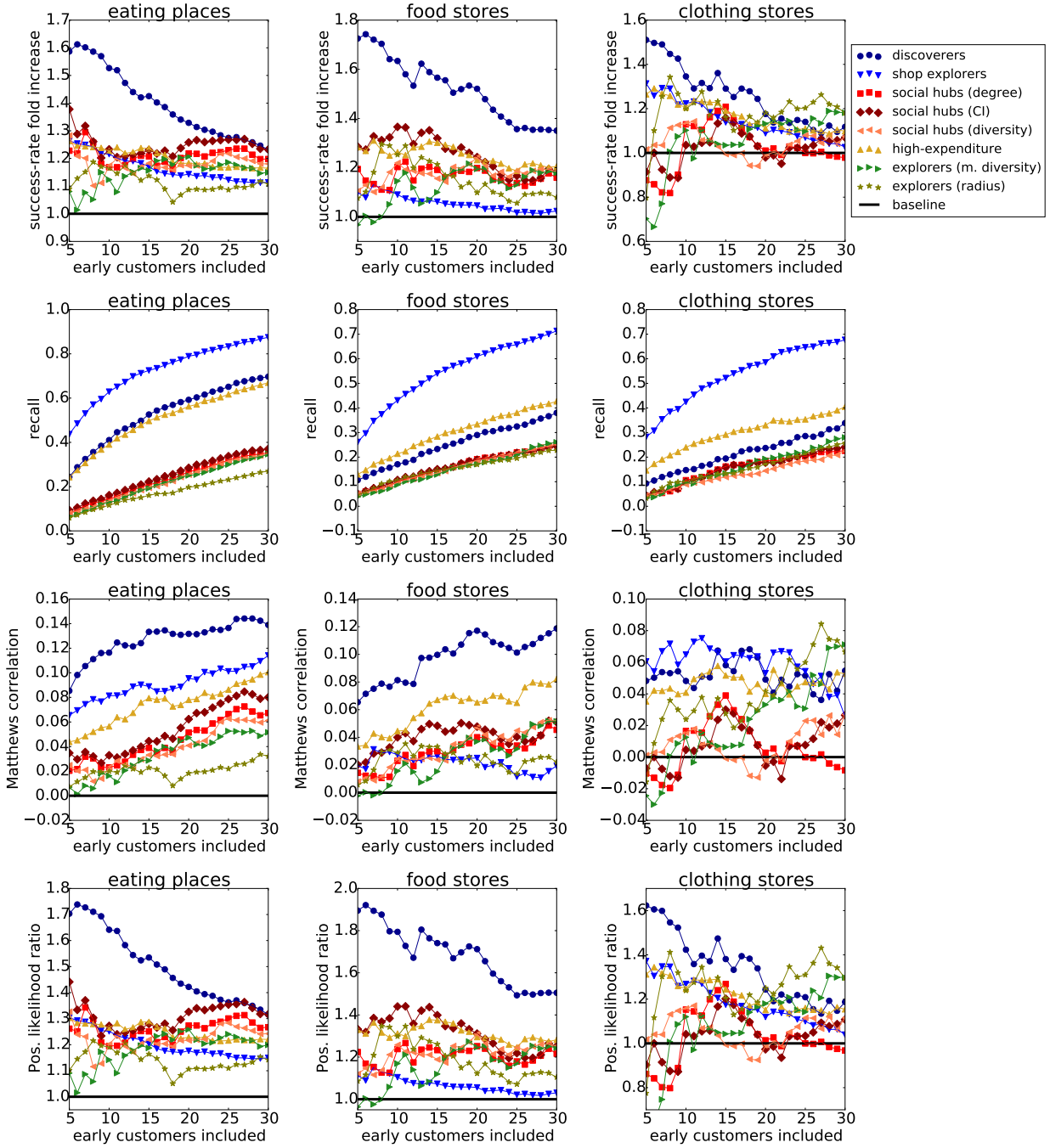

FIG. S14. **Prediction results for a longer training period (from Dec. 2015 to July 2017, 20 months) and a longer validation period (from Aug. 2017 to May 2018, 10 months) compared to those used in the main text**. The other parameters of the analysis are the same as in the main text: $\Delta = 90, z\% = 10\%, s\% = 10\%, \tau\% = 1\%$.

Beyond classification: fold increase of the number of customers

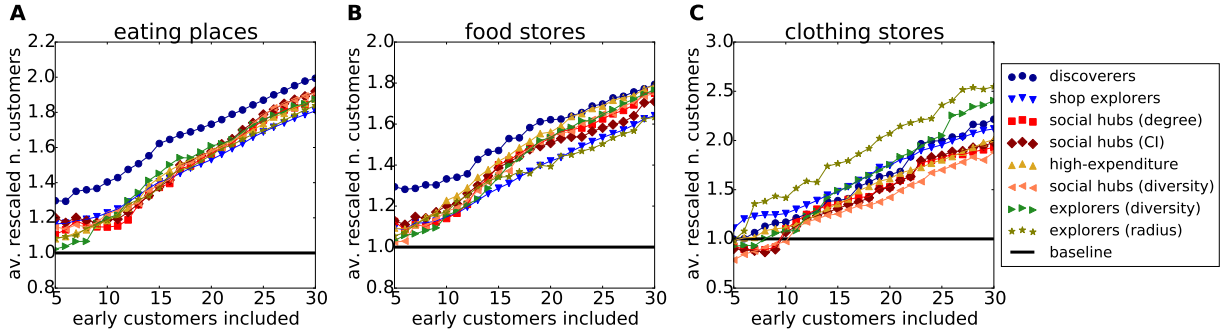

FIG. S15. **Fold increase of the number of customers associated with early purchases by different groups of individuals**. For stores that received their first transaction within the validation period (excluding the last two months), we measure the ratio between their final number of first-time customers, v_i, and the average number of first-time customers of stores with the same MCC introduced in the same month. We refer to this ratio as the rescaled number of customers, $R(v)$. For each group of individuals, $\mathcal{I}$, we measure the average rescaled number of customers of stores that received an individual in $\mathcal{I}$ among the earliest w customers. This average can be interpreted as the fold increase of the number of customers associated with early purchases by individuals in $\mathcal{I}$, compared with stores of the same age and category. We find that the discoverers exhibit the largest fold increases for eating places and food stores, whereas explorers by radius of gyration exhibit the largest fold increase for clothing stores. The parameters for the discoverer training are the same used in the main text: $\Delta = 90, z\% = 10\%, \tau\% = 1\%$.