Interpretable Stochastic Block Influence Model: Measuring Social Influence Among Homophilous Communities

Yan Leng, Tara Sowrirajan, Yujia Zhai, and Alex Pentland

Abstract—Decision-making on networks can be explained by both homophily and social influences. While homophily drives the formation of communities with similar characteristics, social influences occur both within and between communities. Social influences can be reasoned through role theory, which indicates that the influences among individuals depending on their roles and the behavior of interest. To operationalize these social science theories, we empirically identify the homophilous communities and use the community structures to capture such “roles,” affecting particular decision-making processes. We propose a generative model named the Stochastic Block Influence Model and jointly analyze both network formation and behavioral influences within and between different empirically-identified communities. To evaluate the performance and demonstrate the interpretability of our method, we study the adoption decisions for a microfinance product in Indian villages. We show that although individuals tend to form links within communities, there are strongly positive and negative social influences between communities, supporting the weak ties theory. Moreover, communities with shared characteristics are associated with positive influences. In contrast, communities that do not overlap are associated with negative influences. Our framework facilitates the quantification of the influences underlying decision communities and is thus a helpful tool for driving information diffusion, viral marketing, and technology adoption.

Index Terms—Social influence, homophily, stochastic block model, community structure, generative model.

I. INTRODUCTION

We are living in an increasingly connected society [1], [2], [3]. The connections among individuals facilitate information diffusion and enable inter-dependencies in decisions among peers. Therefore, understanding and modeling how hidden social influences change individuals’ decisions are essential and critical to many practical applications, such as viral marketing, political campaigns, and large-scale behavioral change [4], [5].

Homophily, the tendency of similar individuals to associate with each other, widely exists in various types of social networks and controls the outcomes of many critical network-based phenomena [6]. Salient features for homophily come from a wide range of sources, including age, race, socioeconomic status, occupation, and gender [6]. The complex nature of social relationships and the high-dimensional characteristics of individuals thus determine the multi-dimensionality of homophily [7]. Homophily can lead to locally clustered communities and may affect network dynamics, such as information diffusion and product adoption. The block model has been applied to low-dimensional, pre-defined homophilous features and provides a building block to uncover underlying community structures1 with high-dimensional homophily empirically [8].

Social influences are widely studied in both economics and computer science literature due to their importance in understanding human behavior. In economics, researchers focus on causally disentangling social influences from homophily with randomization strategies, such as propensity score matching, behavioral matching and regression adjustment. In the realm of computer science researchers strive to optimize the likelihood of the diffusion path of influences by proposing different generative processes. These works focus on the strength or the pathways of social influence, and they do not link social influences to the underlying homophilous communities and the network formation process.

Two theories shed light on how local communities affect information diffusion and contagion in decision-making. On the one hand, homophily and the requirement of social reinforcement for behavioral adoption in complex contagion theory indicate that influences tend to be localized within homophilous communities [6], [9]. In other words, behavioral diffusion and network formation are endogenous, contributing to the phenomenon of within-community spreading [10]. On the other hand, the weak ties theory [11] implies that bridging ties between communities facilitate the spread of novel ideas. Empirical evidence demonstrates that reinforcement from multiple communities, rather than from the same communities, predicts a higher adoption

Manuscript received 2 June 2021; revised 18 January 2023; accepted 11 June 2023. Date of publication 27 June 2023; date of current version 11 January 2024. The work of Yan Leng was supported in part by the NSF under Grant IIS-2023468. Recommended for acceptance by G. Chen. (Corresponding author: Yan Leng.)

Yan Leng is with the McCombs School of Business, Austin, TX 78705 USA, with the University of Texas at Austin, Austin, TX 78712 USA, and also with the MIT Media Lab, Cambridge, MA 02139 USA (e-mail: yan.leng@utexas.edu).

Tara Sowrirajan is with the MIT Media Lab, Cambridge, MA 02139 USA, also with the Harvard University, Cambridge, MA 02139 USA, and also with the Kellogg School of Management, Evanston, IL 60208 USA (e-mail: stara@mit.edu).

Yujia Zhai is with the Tianjin Normal University, Tianjin 300074, China (e-mail: yjzhai@tjnu.edu.cn).

Alex Pentland is with MIT Media Lab, Cambridge, MA 02139 USA (e-mail: pentland@mit.edu).

This article has supplementary downloadable material available at https://doi.org/10.1109/TKDE.2023.3289848, provided by the authors. Digital Object Identifier 10.1109/TKDE.2023.3289848

1In this paper, we use community and block interchangeably.
rate [12]. Motivated by these contrasting theories, our objective is to investigate whether social influences primarily spread locally within each homophilous community or extend globally to other communities by taking advantage of weak ties.

According to the role theory, “the division of labor in society takes the form of interaction among heterogeneous specialized positions” [13]. That is to say, depending on the social roles and the behavior of interest, the underlying interactions and norms for decision-making are different. Motivated by this proposition, we develop a method that associates social influences with the underlying communities tied to the behavior of interest. To formalize this idea, we propose a generative model to investigate how social influences impact decision-making by inferring the spread of influences across empirically-identified blocks. Using our framework, we uncover the underlying blocks and infer two types of relationships across these blocks: social interaction and social influence. Unlike the Stochastic Block Model (SBM), the observed individual decisions are used to inform the communities, complementing the observed network. In addition, we infer an influences matrix consisting of the social influences across different communities. This influences matrix reveals hidden social influences at the community level, which would otherwise be impossible to observe and generalize.

To illustrate our approach, we experiment on the diffusion of microfinance in Indian villages and perform extensive analysis on the influences matrices estimated from our model. Even though social relationships are denser within communities, social influences predominantly spread across communities. This observation aligns with the importance of across community weak ties [11], and the strength of structural diversity [12]. Our generative framework and subsequent understanding of how social influences operate have practical applications, such as viral marketing, political campaigns, and large-scale health-related behavioral change.

Our paper makes the following contributions to the literature:

1. The Stochastic Block Influence Model (SBIM) integrates networks, decisions, and user characteristics into the generative process. It jointly infers two types of relationships among empirically-identified: social connections and social influences. Unlike traditional SBMs, our approach can flexibly accommodate positive and negative social influences.
2. We adopt a role theory-inspired approach in our SBIM, recognizing that individuals make decisions depending on the context of the decision type [13] (e.g., adopting microfinance as opposed to adopting healthy habits). To achieve this, we allow the underlying community to vary with the behavior of interest.
3. We perform a case study analyzing the adoption of microfinance products in Indian villages. We demonstrate the interpretability of our SBIM through detailed analyses of the influence structure.

II. RELATED LITERATURE

There are two prominent theories for explaining social influence propagation: simple contagion and complex contagion. According to simple contagion theory, individuals will adopt the behavior as long as they have been exposed to the information [11], which is a sensible model for epidemics and information spreading. Complex contagion theory, on the other hand, requires social reinforcement from neighbors to trigger adoption [9]. Many studies have shown that complex contagion explains behaviors such as registration for health-related forums [14]. However, these exposure-based models are analytically simple and do not allow social influences to be negative, i.e., the adoption decision of one’s neighbors might decrease, rather than increase, the likelihood of one’s adoption decision. Moreover, they typically are not able to capture the heterogeneity of social influences [5]. In this paper, we propose a model to account for negative and heterogeneous influences.

The SBM is a statistical model for studying latent cluster structures in the network data [8]. The SBM generalizes the Erdos-Renyi random graph model with higher intra-cluster and lower inter-cluster probabilities. The traditional SBM only infers community structures from network connections. However, when contextual information on nodes is available, leveraging information from different sources facilitates inference. There has been interesting work on utilizing covariates to infer underlying block structures in recent statistics literature. For example, Binkiewicz et al. [15] present a covariate-regularized community detection method to find highly connected communities with relatively homogeneous covariates. They balance the two objectives (i.e., the node covariance matrix and the regularized graph laplacian) with tuned hyper-parameters. Yan et al. [16] propose a penalized optimization framework by adding a k-means type regularization. This framework is based on the premise that the estimated communities are consistent with latent memberships in the covariate space. Although these variations to the SBM utilize auxiliary information on individual nodes, they specify the importance of recovering the network and the smoothness of covariates on the network on an ad hoc basis. Different from these models, we take advantage of the role theory [13] and utilize the decision-making processes on the network that could also inform community detection. More importantly, the communities that we discover are specifically relevant for the decision-making of interest, while the ones discovered by SBM is only affected by the network connects (and is agnostic to decision-making). For example, professional communities are more useful for the adoption of technologies at work, while social communities are more useful for the adoption of social mobile applications. The underlying communities depend on the roles and behaviors of interest because social influences spread through network links, depending on different applications. SBIM bridges the rich SBM and social contagion literature. It opens up future opportunities to adapt to other variations of SBM.

III. METHODOLOGY

A. Stochastic Block Influence Model

Assume a random graph $G(V, E)$ with $N$ individuals in node set $V$ and edge set $E$. It is partitioned into $C$ disjoint blocks $(V_1, \ldots, V_C)$, and the proportion of nodes in each block $c$ is $\pi_c$, and $\sum_{c=1}^{C} \pi_c = 1$. $A \in \mathbb{R}^{N \times N}$ represents the adjacency matrix.
\( A_{ij} = 1 \) if \( i \) and \( j \) are connected, and \( A_{ij} = 0 \) otherwise. Let matrix \( B \in \mathbb{R}^{C \times C} \) denote block-to-block connection likelihood. Let \( M \) be the block assignment of individual \( i \), and by summing over all \( C \) blocks, we have \( \sum_{k=1}^{C} M_{ik} = 1 \). We combine the block vector of all individuals in the matrix \( M \in \mathbb{R}^{N \times C} \).

Therefore, the probability of a link between \( v_i \) and \( v_j \) between two separate blocks \( V_k \) and \( V_l \) as \( P((v_i, v_j) \in \mathcal{E}| v_i \in V_k, v_j \in V_l) = p_{ij} \), \( y \in \mathbb{R}^N \) is a binary vector representing individuals’ adoption behaviors. Let \( X_i \in \mathbb{R}^D \) represent demographic features, where \( D \) is the number of covariates. We use \( F \in \mathbb{R}^{C \times C} \) to represent the block-to-block influences matrix. Finally, \( h \) is a binary vector, that captures whether each individual is aware of the product at the beginning of the observation period. For a new product, \( h \) is sparse; while for a mature product, \( h \) is dense.

In SBIM, we link latent communities to adoption decisions and socio-demographics. SBIM reveals the underlying nature of high-dimensional homophily in a data-driven fashion rather than using pre-defined communities using observed sociodemographics (e.g., race or occupation). Solely using pre-defined homophilous characteristics does not aptly capture the multifaceted characteristics that define individuals and their social ties. In other words, individuals are associated with different communities, each formed by various homophilous characteristics. Neighbors belonging to different communities may influence focal individuals differently.

We illustrate this using the adoption of microfinance in Indian villages. We posit that several traits define the diverse nature of individuals: different professions, castes, education levels, and a variety of other demographic features. One individual, an educated worker of a lower social caste, belongs with varying degrees of affiliation to different communities. The individual is perhaps most strongly affiliated with a group that has a certain level of education and less strongly affiliated with another group where most of the group are of a lower caste. This mixed membership captures the realistic nature of our social relationships and characteristics. Within such a village with multi-dimensional homophily, how can we understand who influences this individual and what processes are involved in that individual’s decision-making? Specifically, this individual could be influenced by neighbors belonging to different communities characterized by specific educational backgrounds, professions, and castes. The data-driven multi-dimensional blocks of the model allows us to capture these critical, hidden relationships.

Next, we formalize our SBIM model. To jointly infer how influences spread within and across communities, we create a model with the following properties:

1) The model leverages both the observed friendship network structure and adoption behavior to infer the underlying communities.

2) The link formation and social influences between two individuals are jointly determined by their underlying communities.

**Blocks for Network Connections:** For each individual pair \( \{i, j\} \), depending on their community assignment vectors, the predicted link \( \hat{A}_{ij} \) is generated according to the connection probability matrix, \( B \). In particular, the probability of the existence of a link between \( i \) and \( j \) is

\[
\mathbb{P}(\hat{A}_{ij} = 1| M, B) = (MBM^T)_{ij}.
\]

**Blocks for Decision-Making:** Next, we discuss how we incorporate individual characteristics and adoption decisions into our SBIM. The adoption likelihood depends on individuals’ characteristics and the influences of their neighbors who have already adopted [17]. This generative model builds upon the communities an individual \( i \) and \( i’ \)’s neighbors belong to, as well as the community-to-community influences matrix \( F_{ij} \). Each individual decides on whether or not to adopt to maximize her utility. The utility of \( i \) depends on what this individual prefers and the aggregated influences from neighbors. The pairwise influences depend on the communities that \( i \) and her neighbors belong to. We illustrate how influences and communities affect one’s decision-making in Fig. 1. We consider individual \( A \), who has three friends, \( B, C, \) and \( D \), belonging to a lower socioeconomic status (SES) group (as colored in red), and one friend, \( E \), belonging to a higher SES group (as colored in blue).

The adoption likelihood of individual \( A \) is a function of this individual’s preferences as well as the influences of friends \( B, C, D, \) and \( E \). The strength of the influences depends on the corresponding communities of \( A \) and her friends.

More generally, a user’s adoption likelihood, \( \hat{y}_i \), is defined as

\[
\hat{y}_i = \logit(\beta X_i + \sum_j (\{MFM^T\} \circ (\{h \otimes 1 \) \circ A))_{ji} + \epsilon_i),
\]

where \( \circ \) represents element-wise matrix multiplication. The first term, \( \beta X_i \), measures the adoption decision conditioned on \( i \)'s sociodemographic features if there were no social influence, where \( \beta \in \mathbb{R}^D \) and \( D \) is the dimension of the covariates. The second term aggregates the influences of \( i \)'s neighbors. \( \epsilon_i \) is an idiosyncratic error term. Without loss of generality, we assume \( \epsilon_i \sim \mathcal{N}(0, 1) \).
For a mature product that everyone knows, we can simplify (2) as:
\[
\hat{y}_i = \logit \left( \beta X_i + \sum_{j=1}^{N} \left( (MF^{T}) \circ A \right)_{ji} + \epsilon_i \right). \quad (3)
\]

Equation (2) only accounts for the influences among direct neighbors. Note that in a small-scale network, it is reasonable to assume higher-order social influences do not exist. In a large-scale network, Leng et al. [18] show that social influences spread beyond immediate neighbors. Our model can be easily adapted to higher-order influences by summing up the powers of the adjacency matrix \( A \) to account for multiple degrees of separation, as done in [5].

We finally introduce the loss function (\( L \))
\[
L = -\sum y \log(\hat{y}) - \sum A \log(\hat{A}). \quad (4)
\]

The first component is the main objective for typically SBM and the second component is SBIM’s main advantage. The same inferred blocks \( M \) are used to compute both \( \hat{A} (1) \) and \( \hat{y} (2) \) and (3). In the loss function, we minimize the difference between (1) predicted and observed links (the objective for existing SBM in the “Blocks for network connections” section); (2) predicted and observed behaviors (contribution of our paper discussed in “Blocks for decision-making”).

### B. Generative Process

For the full network, we assume the following generative process in the model, which defines a joint probability distribution over \( N \) individuals, based on the node-wise membership matrix \( M \), block-to-block interaction matrix \( B \), block-to-block influences matrix \( F \), attributes’ coefficients \( \beta \), observed friendship network \( A \), observed attributes \( X \), and observed adoption decision \( y \). For brevity, we denote \( \mathcal{Z} \) as set of the hidden variables, \( \mathcal{Z} = \{ M, \beta, B, F \} \) and \( \theta \) as the set of hyperparameters, where \( \theta = \{ c, a, b, \mu_a, \sigma_a, \mu_F, \sigma_F, \mu_b, \sigma_b \} \).

1. For each node \( v_i \in \mathcal{V} \), draw a \( C \)-dimensional mixed membership vector \( M_i \sim \text{Dirichlet}(c) \).
2. For the connection probability from community \( k \) to \( l \) in the block-to-block connectivity matrix, draw \( B_{kl} \sim \text{Beta}(a, b) \).
3. For the influences from community \( k \) to \( l \) in the block-to-block influence matrix, draw \( F_{kl} \sim \mathcal{N}(\mu_F, \sigma_F) \).
4. For each attribute in \( \beta_j \), draw the coefficient \( \beta_j \sim \mathcal{N}(\mu_j, \sigma_j) \).
5. Draw connections between each pair of nodes \( v_i \) and \( v_j \), \( A_{ij} \), according to (1).
6. Draw adoption decisions \( \hat{y}_i \), according to (2).

Steps (3), (4), and (6) are unique processes of SBIM, which are relevant to the adoption behavior of our method. The blocks used in SBM in Step (1) are also used in the generation process in (2)–(3) in Step (6). As discussed in Section III, the latent component \( M \) contributes both to \( \hat{A} (1) \) and \( \hat{y} (2) \) and (3), as it affects both the decision-making and the link formation.

The posterior distribution defined by SBIM is a conditional distribution of the hidden block structure and relationships given the observed friendship network and adoption behavior, which decomposes the agents into \( C \) overlapping blocks. The posterior will place a higher probability on configurations of the community membership that describe densely connected communities and stronger (positive or negative) influences.

The posterior of the SBIM is intractable, similar to many hierarchical Bayesian models. Therefore, we use the Markov Chain Monte Carlo (MCMC) algorithm as an approximate statistical inference method to estimate the parameters. The MCMC draws correlogated samples that converge to the target distribution and are generally asymptotically unbiased. MCMC methods include Gibbs sampling, Metropolis-Hastings, Hamiltonian Monte Carlo (HMC), and No-U-Turn Sampler (NUTS). Gibbs sampling and Metropolis-Hastings methods converge slowly to the target distribution as they explore the parameter space by random walk. HMC suppresses the random walk behaviors with an auxiliary variable that transforms the problem by sampling to a target distribution using simulated Hamiltonian dynamics. However, HMC requires the gradient of the log-posterior, which is complicated in our model. Moreover, it requires a reasonable specification of the step size and several steps, which would otherwise result in a substantial drop in efficiency [19]. Therefore, we apply NUTS, a variant of the HMC method, to eliminate the need for choosing the number of steps, by automatically adapting the step size. Specifically, the NUTS builds a set of candidate points that spans the target distribution recursively and automatically stops when it starts to double back and retrace its steps [19].

### IV. Experiments

We study the adoption of microfinance in five villages in India collected by the Abdul Latif Jameel Poverty Action Lab (J-PAL) [20]. In 2007, a microfinance institution introduced a microfinance program to some selected villages in India. In early 2011, they collected information about whether or not the villagers had adopted the microfinance product. Because the villages are fairly small and microfinance had been on the market for four years when J-PAL collected individuals’ adoption decisions, it is reasonable to assume that everyone in the village was aware of microfinance, which is hence a considered mature product in this village. Therefore, we employ (3) as the decision-making function. The data contains information about self-reported relationships among households and other amenities, including village size, quality of access to electricity, quality of latrines, the number of beds, the number of rooms, the number of beds per capita, and the number of rooms per capita. These demographic features serve as the independent variables. The outcome variable is the adoption decision of microfinance. The microfinance institution asked the villagers to self-report other villagers they considered as friends.

To train and evaluate our model for a specific number of blocks, we performed cross-validation by randomly splitting the data into 75% training samples and 25% test samples. We repeated this process ten times. With NUTS, we obtained the point estimates for all latent variables in \( \mathcal{Z} \). We subsequently reran our model (as described earlier), fixing all latent variables.
to the estimates obtained from the test dataset. This step returns the predicted adoption probability for each villager in the test data. To choose the optimal number of blocks, we first tune the model for $C \in \{2, 6, 10, 14\}$ and then calculate the average loss. We observe a negative parabolic trend with the loss peaking at its lowest at $C = 10$ blocks, so we use this optimal number of blocks in additional analyses. We use a machine learning model with sociodemographics and the hidden community learned by spectral clustering\(^2\) (i.e., blocks) on the adjacency matrix as the independent variables. In this way, we use the same information in SBIM and the benchmarks.

Since the dependent variable in our data is imbalanced, we evaluate our SBIM using the area under the Receiver Operating Characteristics curve (AUC) that we plot using the false positive rate and correct positive rate for different thresholds. We define a loss metric to select the best configurations during the training period. We formulate it by taking the negative of the standard improvement measure, which is the absolute improvement in performance normalized by the room for improvement. This measure captures the improvement of our SBIM compared to the baseline model. Since we have a small test set, making predictions on a randomly-drawn test set is hard. Measuring the relative improvement ensures that the composition of the test set does not bias the performance due to sample variation. This metric is formulated by: $L = \frac{1 - \text{Baseline test AUC} - \text{SBIM test AUC}^+}{\text{Baseline test AUC} - \text{Baseline bas AUC}}$, where the AUC of the baseline model and SBIM on the test split in cross-validation are represented as Baseline test AUC and SBIM test AUC\(^+\), respectively.

SBIM has seven hyperparameters in $\theta$. Since the parameter space is large, we adapt a bandit-based approach to tune the parameters developed called hyperband [22]. Our adaptation of this algorithm allows each configuration tested to run with full resources due to our sampling procedure, allowing NUTS to run consistently across all configurations.

We compare the performance of our SBIM model with six methods benchmarks, all of which use the learned block as the extra feature, in Table I. Our SBIM outperforms the best benchmark (elastic net with blocks) in the test set by 13.4% according to the aforementioned improvement metric.

V. ANALYSIS AND DISCUSSIONS

Interpretability is a broad term in machine learning. We follow the definition of [23]. We define interpretable machine learning as the extraction of relevant knowledge from a machine learning model concerning relationships either contained in the data or learned by the model. Our SBIM model satisfies several characteristics of the model-based interpretability methods developed in this paper, including sparsity, simulability, and modularity. We discuss how our model satisfies each criterion in the Appendix, available online.

We can associate individuals’ sociodemographics with the members of each block, allowing us to generalize block types based on characteristics such as high or low SES, homogeneous or diverse, and skill levels, as depicted in Table II (remaining examples are shown in Table A1 in the Appendix), available online. In this example, each block is associated with a qualitative type, and the attributes within that block describe such characterizations. Caste composition, education levels, and profession types are employed to designate lower or higher SES blocks. Homogeneous or diverse blocks are designated by some professional composition, caste types, native language composition, gender imbalance, and what fraction of village inhabitants are natives.

We use normalized entropy to measure the diversity of attributes studied in this paper. Normalized entropy is a metric used to capture the number of types of characteristics within each category while accounting for the frequency of each entity type within a category. It can be formulated by, $Q = -\frac{1}{n} \sum_{i=1}^{n} p_i \log(p_i)$, where $q$ denotes the number of types within a category, $p_i$ denotes the probability of each type $i$, $f$ denotes the number of each type $n_i$.

The gender ratio ($R$) is measured within a block and is formulated by $R = \frac{m}{f}$, where $m$ and $f$ denote the number of occurrences of males and females respectively. Thus, since $R$ denotes the ratio of males to females in a block, both a high or low gender ratio corresponds to a high gender imbalance.

The total influences into and out of each block are depicted in Fig. 2, which allows us to evaluate the aggregated influences a block receives and spreads (net positive, negative, or neutral). For example, we can see diverse, low-SES block five and senior, low-SES block six have high output levels of positive influence, and diverse, middle-SES block eight receives a net high level of negative influence. We find that some blocks have stronger outgoing influences than others and can perceive these as influential leaders of positive and negative influences. Similar reasoning applies to characterize blocks that receive a high level of influence as follower blocks. We also observe the difference in net incoming and outgoing influences within each block

\(^2\) Spectral clustering uses the second smallest eigenvector of the graph laplacian as the semi-optimal partition [21].

---

**TABLE I**  
**MODEL AND BASELINE PERFORMANCE**

| Method                  | Train Mean (S.D.) | Test Mean (S.D.) |
|-------------------------|-------------------|------------------|
| Random forest (with blocks) | 0.503 (0.010)     | 0.610 (0.095)    |
| Gradient boosting (with blocks) | 0.443 (0.075)     | 0.531 (0.058)    |
| Adaboost (with blocks)    | 0.875 (0.038)     | 0.528 (0.058)    |
| Elastic net (with blocks) | 0.724 (0.107)     | 0.612 (0.079)    |
| Lasso regression (with blocks) | 0.719 (0.069)     | 0.607 (0.078)    |
| Multilayer perceptron (with blocks) | 0.703 (0.096)     | 0.536 (0.056)    |
| SBIM                     | 0.805 (0.022)     | 0.664 (0.062)    |

**TABLE II**  
**BLOCK CHARACTERISTICS EXAMPLE. SES SHORTS FOR SOCIOECONOMIC STATUS**

| Block Type | Attributes |
|------------|------------|
| Heterogeneous, low-SES | Only use disadvantaged caste and rural language spoken |
| Low productivity and education levels |
| Diverse, skilled, highly-educated | Low gender ratio |
| Minority languages and diverse, high-skilled professions |
| Multi-skilled caste |
| Richer composition |
| Lower education |
| Young, low-SES | Minority low skill blue profession in agriculture |
| Minority older age, male dominated blocks |
| Minority lower caste, semi-skilled positions |
| Richer composition |
| Diverse, low-SES | Minority language dominance, moderate education |

Higher education refers to having education levels at PUC (post-university course) and having a “degree or above” designation.
the direction of negative influence is most frequently observed from a low-SES block to a high-SES block. Additionally, we frequently observe positive self-influence, which is from a block to itself, and this occurs when a block is characterized by a younger average age, highly-educated, high job diversity, higher-skilled jobs, high language diversity, large gender imbalance, and having a large number of village natives. The remaining examples are displayed in Table A2 available online.

When we analyze several examples using block characteristics composition and observing the types and patterns of influence, several general trends arise, as depicted in Table A2 available online. The block attributes most frequently associated with different influences are summarized into key trends. We find a positive influence when two blocks overlap in the following characteristics: gender distribution, majority castes, professions, high profession diversity, highly educated, highly-skilled jobs, and native languages. Negative influences frequently occur when two blocks do not overlap in the following characteristics: gender distribution, caste composition, profession diversity level, education levels, and average age. Furthermore, negative influences are most frequently observed from a low-SES block to a high-SES block. Additionally, we frequently observe positive self-influence, which is from a block to itself, and this occurs when a block is characterized by a younger average age, highly-educated, high job diversity, higher-skilled jobs, high language diversity, large gender imbalance, and having a large number of village natives.

When paired with block type characterizations, we find that these trends lead to interesting associations, such as block-to-block perceptions of lower or higher SES groups having influence. Blocks with the higher SES group designation more frequently receive negative influences from lower-SES blocks. Blocks of similar SES, especially higher SES, have more frequent positive influences, and High-SES blocks also have more frequent positive self-influence.

These findings suggest that firms should employ marketing strategies that take into account the characteristics of the underlying communities. For example, the microfinance institution could organize separate information sessions for the high-SES and low-SES groups to take advantage of the positive influences between groups that share similar characteristics while avoiding negative influences across different communities. Moreover, suppose the microfinance institution is to introduce the product into other villages (as a new product), they should send the information to individuals with the following characteristics: (1) high-SES with fewer low-SES neighbors, (2) individuals who speak a diverse set of languages; and (3) communities with similar gender ratios.

VI. CONCLUSION

According to the role theory, the interactions of individuals depend on their roles and behaviors of interest. To conceptualize this idea, we use the underlying community structures to capture the “roles,” which affect the particular decision-making processes of individuals. Specifically, we develop the Stochastic Block influences Model, which infers two types of hidden relationships: (1) block-to-block interactions, and (2) block-to-block

![Fig. 2. Net influences into and out of each block.](Image)

![Fig. 3. Social influences across social blocks (directed links) for gender ratio and profession diversity (node color).](Image)

### Table III

| Attribute | Positive Influence | Negative Influence | Positive self-influence |
|-----------|--------------------|--------------------|-------------------------|
| Gender    | Socio-economic distribution | Social-economic block to negative influence | Large gender imbalance |
| Caste     | Overlapping majority caste | Lack of overlap in caste composition | Majority village native |
| Profession| Professional overlap in specialty jobs especially large profession diversity | Professional overlap in specialty jobs especially large profession diversity | Higher skilled jobs |

By analyzing several examples in this manner using block characteristic composition and observing the types and patterns of influence, several general trends arise, as depicted in Table III. The block attributes most frequently associated with different types of influence are summarized into key trends. Positive influence occurs when two blocks overlap in the following characteristics: gender distribution, majority castes and professions. Negative influence frequently occurs when two blocks have a lack of overlap in the following characteristics: gender distribution, caste composition, and profession diversity level. Furthermore,
influences on decision-making. Moreover, our model flexibly allows for both positive and negative social influences. The latter is more common in practice but has been largely ignored in the contagion models in the literature [9], [20]. In the adoption of microfinance examples we present, the inferred block-to-block influences analysis offers insights into how different social blocks influence individuals’ decision-making. Our framework has far-reaching practical impacts to understand the patterns of influences across communities and identifying the crucial characteristics of influential individuals for several applications. First, practitioners and researchers can identify the most influential communities (e.g., leaders and followers) and understand the dynamics among different communities that are not available or observable without our SBIM model. Second, marketers can investigate which sociodemographics predict positive or negative social influences and utilize this information when introducing a product to a new market. Lastly, marketing firms can use the influences of each individual to decide whom to target for campaigns [5].

Our SBIM is not without limitations and opens up several directions for future studies. First, future research can easily adapt the SBIM to accommodate a more complicated stochastic block model, such as a degree-corrected SBM or a power-law regularized SBM. Second, a scalable inference method as an alternative to NUTS sampling will help to improve the efficiency and scalability of SBIM. Third, future research can extend the SBIM model to a dynamic model and consider its application for new products, where the influences matrix varies with time and distances from the source of information, as well as considering the diffusion rate. Lastly, for computer scientists and social scientists who have access to similar types of data but in diverse contexts (e.g., different behaviors collected in different regions), it will be interesting to apply and compare the influence matrices to see if there exists any generalizable pattern to support the contagion and decision-making theories.

REFERENCES

[1] J. Travers and S. Milgram, “An experimental study of the small world problem,” Social Netw., Academic Press, 1977, pp. 179–197.
[2] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, “Four degrees of separation,” in Proc. 4th Annu. ACM Web Sci. Conf., 2012, pp. 33–42.
[3] Y. Leng, D. Santistevan, and A. Pentland, “Familiar strangers: The collective regularity in human behaviors,” 2018, arXiv: 1803.08955.
[4] J. H. Fowler and N. A. Christakis, “Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study,” Brmj., vol. 337, 2008, Art. no. a2338.
[5] Y. Leng, Y. Sella, R. Ruiz, and A. Pentland, “Contextual centrality: Going beyond network structure,” Sci. Rep., vol. 10, no. 1, pp. 1–10, 2020.
[6] M. McPherson, L. Smith-Lovin, and J. M. Cook, “Birds of a feather: Homophily in social networks,” Annu. Rev. Sociol., vol. 27, no. 1, pp. 415–444, 2001.
[7] P. Block and T. Grund, “Multidimensional homophily in friendship networks,” Netw. Sci., vol. 2, no. 2, pp. 189–212, 2014.
[8] E. Abbe, “Community detection and stochastic block models: Recent developments,” 2017, arXiv: 1703.10146.
[9] D. Centola and M. Macy, “Complex contagions and the weakness of long ties,” Amer. J. Sociol., vol. 113, no. 3, pp. 702–734, 2007.
[10] P. Pin and B. W. Rogers, “Stochastic network formation and homophily,” in The Oxford Handbook of the Economics of Networks, Oxford University Press, 2016, pp. 138–166.
[11] M. S. Granovetter, “The strength of weak ties,” in Social Networks, Amsterdam, The Netherlands: Elsevier, 1973, pp. 347–367.
[12] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, “Structural diversity in social contagion,” Proc. Nat. Acad. Sci. USA, vol. 109, pp. 5962–5966, 2012.
[13] B. J. Biddle, “Recent developments in role theory,” Annu. Rev. Sociol., vol. 12, no. 1, pp. 67–92, 1986.
[14] D. Centola, “The spread of behavior in an online social network experiment,” Science, vol. 329, no. 5996, pp. 1194–1197, 2010. [Online]. Available: http://science.sciencemag.org/content/329/5996/1194
[15] N. Binkiewicz, J. T. Vogelstein, and K. Rohe, “Covariate-assisted spectral clustering,” Biometrika, vol. 104, no. 2, pp. 361–377, 2017.
[16] B. Yan and P. Sarkar, “Covariate-assisted spectral clustering in sparse graphs,” J. Amer. Stat. Assoc., vol. 116, pp. 734–745, 2021.
[17] Z. Katona, P. P. Zubcsek, and M. Sarvary, “Network effects and personal influences: The diffusion of an online social network,” J. Marketing Res., vol. 48, no. 3, pp. 425–443, 2011.
[18] Y. Leng, D. Xiaodong, M. Esteban, and P. Alex, “Long-range social influence in phone communication networks on offline adoption decisions,” Inf. Syst. Res., 2023.
[19] M. D. Hoffman and A. Gelman, “The No-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo,” J. Mach. Learn. Res., vol. 15, no. 1, pp. 1593–1623, 2014.
[20] A. Banerjee, A. G. Chandralekhar, E. Duflo, and M. O. Jackson, “The diffusion of microfinance,” Science, vol. 341, no. 6144, 2013, Art. no. 1236498.
[21] A. Y. Ng, M. I. Jordan, and Y. Weiss, “On spectral clustering: Analysis and an algorithm,” in Proc. Adv. Neural Inf. Process. Syst., 2002, pp. 849–856.
[22] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, “Hyperband: A novel bandit-based approach to hyperparameter optimization,” J. Mach. Learn. Res., vol. 18, no. 1, pp. 6765–6816, 2017.
[23] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbas-Ali, and B. Yu, “Definitions, methods, and applications in interpretable machine learning,” Proc. Nat. Acad. Sci. USA, vol. 116, no. 44, pp. 22071–22080, 2019.