Learning from Like-Minded People

Delong Meng

September 5, 2017

Abstract

We study a social learning model in which people choose who to talk to and strategically exchange information. Agents start with heterogeneous priors about an unknown state of the world. First each agent chooses a partner. Then everyone observes a private i.i.d. signal and sends a message to her partner. Finally everyone takes an action based on her prior, her private signal, and her partner’s message. Our main finding is that when the signal space and action space are binary, assortative matching arises in equilibrium, but it is generally inefficient for social welfare and information aggregation. In addition we construct counter-examples (non-assortative matching) in the case of multiple signals or multiple actions.

Keywords: assortative matching, homophily, Bayesian persuasion, heterogeneous priors, economics of religion

JEL Classification: D83, D85, Z12

This work is funded by the John M. Olin Foundation. We thank Douglas Bernheim, Alex Bloedel, Gabriel Carroll, Glenn Ellison, Matthew Gentzkow, Matthew Jackson, Fuhito Kojima, Scott Kominers, Yucheng Liang, Michael Ostrovsky, Harry Di Pei, Alvin Roth, Alex Wolitzky, Qingyun Wu, and Muhamet Yildiz for helpful discussions.

1 Introduction

Motivation  Homophily slows down the diffusion of information and decreases the rate for society to reach consensus (Golub and Jackson (2012)). However in reality people often choose to learn from like-minded people. Religious participation is a prime example;
people discuss their religious beliefs and moral values with others in church, who often share the same set of beliefs, and these beliefs significantly impact their lifestyle and voting behavior. Political discussion is another example. Gentzkow and Shapiro (2011) find people are less segregated (ideologically) on the internet, but more segregated in person, especially when choosing who to discuss politics with. These behaviors create an echo-chamber effect and contribute to the polarization of beliefs.

Why do people choose to discuss religion and politics with like-minded people? One obvious explanation is that people lose utility from engaging in heated arguments; for example Mutz (2006) describes religion and politics as “conversational minefields”. Another explanation is that people want to learn the truth, but they fall into victims of confirmation bias and seek information from the source they agree with (see Nickerson (1998)). A third approach is rational choice. For example in the context of media bias Gentzkow and Shapiro (2006) study a model in which readers infer that a media outlet is less accurate if its report is contrary to the readers’ prior beliefs.

In this paper we propose a different rational foundation for why people choose to learn from like-minded people. We consider a social learning model in which people choose who to talk to and strategically exchange information. Our main intuition is that a person who has already made up her mind would want to be matched to a pivotal voter, but the pivotal voter – knowing that the other person would strategically release information to persuade her – would rather be matched with someone who is more moderate. The key assumption in our model is that everyone can be matched with exactly one partner, and match must be consensual. In general an agent prefers talking to the other side only if the other side is more susceptible to persuasion – he is more likely to persuade his partner than the other way around – but this incentive cannot hold in both directions. As a result assortative matching arises in equilibrium.

In this paper we build a model to analyze when the argument above gives rise to assortative matching and when it doesn’t. We also explore the consequences of assortative matching from the perspective of social welfare and information aggregation.

**Model and results** In our baseline model there is an unknown binary state of the world. Agents start with heterogeneous priors, and these priors are common knowledge. Since the state is binary, we can parameterize the agents’ priors on the interval $[0, 1]$ according to their prior for state 1. Assume all agents’ priors lie in some interval $[x, 1 - x]$, so that no agent is extreme; every agent wants to acquire signals beyond their prior. We specify the boundary $x$ more precisely in the model.
First each agent obtains a partner via a stable matching process. Agents’ preferences in the matching stage are determined by their payoffs from following communication game. Each agent observes an i.i.d. private signal and sends a message to her partner; we use Bayesian persuasion to model this exchange. Afterward each agent takes an action based on her prior, her private signal, and her partner’s report. An agent’s utility depends on the distance between everyone’s action and the state of the world; she wants everyone to take an action that matches the true state. Since agents have heterogeneous priors, they disagree on the true state, and communication becomes strategic.

Our primary interest is the matching stage: would assortative matching arise in equilibrium? By assortative matching, we mean that partners always disclose full information with each other and take the same action. For example when actions are binary, an agent with prior less than $\frac{1}{2}$ is matched with a partner whose prior is also less than $\frac{1}{2}$.

Here are our main results:

1. When signals and actions are both binary, all stable matchings that satisfy two refinement criteria must be an assortative matching. The refinement criteria rule out cases when partners exchange no information at all, or that stability is highly sensitive to the initial priors. (Theorem 2.1)

2. When signals and actions are both binary, assortative matching is generally inefficient for social welfare and information aggregation. (Theorem 4.2)

3. The above results continue to hold with multiple states of the world. However with multiple signals or multiple actions, we could construct counter-examples where the stable matching is non-assortative. (Sections 5.1 and 7)

The first result formalizes the idea that people choose to talk to like-minded people. The main argument, as aforementioned, is that an agent wants to be matched with the other side only if her partner’s prior is closer to $\frac{1}{2}$ than hers is. The only way for this incentive constraint to hold for both sides is that the partners have priors of equal distance to $\frac{1}{2}$ (e.g. 0.3 and 0.7). The refinement criteria rule out stable matchings like these because their stability is highly sensitive to the priors: if we perturb one of the priors by $\epsilon$, then the matching is no longer stable. On the other hand assortative matching is stable and is robust to small perturbations to the priors. As long as the partners have priors on the same side of $\frac{1}{2}$, they honestly exchange information and take the same action.

The second result says assortative matching is socially inefficient. Since agents have heterogeneous priors, there is no standard definition for efficiency. We propose four pos-
sible objectives for the social planner. First the social planner maximizes the sum of everyone’s subjective utility according to each agent’s own prior. Second the social planner maximizes the sum of everyone’s expected utility based on the social planner’s prior. Third the social planner uses a prior-free approach: a matching dominates another one if for all social planner’s priors, this matching is more efficient under the second criterion. Fourth the social planner has a neutral prior and maximizes the probability that the majority of the agents are taking the correct action. Under criteria 1, 2, and 4, assortative matching is inefficient except in some boundary cases. Under criterion 3 assortative matching is efficient, but this criterion is a weak notion of efficiency: every matching is efficient under this prior-free approach.

The third result shows that binary state is not the main driver for assortative matching. With multiple states as long as the signals satisfy the monotone likelihood ratio property, our argument for assortative matching still goes through. However when there are multiple signals or multiple actions, non-assortative matching could arise. Intuitively, when there are two actions, nobody wants to talk to the agent with prior 0.9, because it is difficult to persuade him from 0.9 to 0.5. If there are multiple action, then persuading this agent from 0.9 to 0.8 might already change his action, so people might be willing to be matched with more extreme agents.

**Applications** Examples of our model must satisfy three conditions. First people not only care about their own action, but also the actions of others. Second when people on different sides talk to each other, they try to persuade each other. Third some people prefer talking to the other side, but these people already have strong priors that they are unlikely to be persuaded by the other side.

**Religion.** Religious proselytizing is a clear example of persuading people on the other side. Moreover people who preach the gospel must be confident about their own faith so that non-believers won’t sway them into unbelief. Similarly militant atheists are vocal about persuading people away from church. Assortative matching means people choose the congregation that agree with their theology, and that atheists typically don’t go to church.

**Politics.** People care about election outcomes and therefore care about other people’s beliefs and votes. Persuasion ranges from participating in large scale campaigns to reaching out to friends and neighbors. Moreover people who join campaigns are the enthusiastic supporters who already know who/what they are voting for. Assortative matching occurs in inter-personal discussions. Gentzkow and Shapiro (2011) find people are more ideologi-
cally segregated offline than online, and isolation index is highest for the choice of political discussants. Similarly Bill Bishop’s book “The Big Sort” claims that people today are geographically sorted with like-minded others. There are also institutional attempts for assortative matching, such as creating a “safe space” on university campuses.

*Environmental activism.* If someone strongly believes that recycling is critical for the environment, then she wants others to recycle. Similarly those who believe in human-caused climate change want people to support policies that curb carbon emission, while those who deny climate change want to people to vote against such policies. Though it’s hard to test empirically how people persuade each other on climate change, the consequence is clear. Democrats and Republicans differ sharply on this question.

*Charity / foreign aid.* Some believe that foreign aid benefits only the leaders of poor countries because of corruption, while others argue that nepotism is a limited problem and could even be welfare improving (Basurto et al. (2017)). Similarly some people give money to the homeless to help them improve their lives, but others believe that homeless people will use their money on drugs. In both scenarios the person’s belief impacts her decision to donate money as well as her preference for other people’s behavior. If she believes donating money is ineffective, then she wants her friends to stop donating money. Likewise if she supports giving money to the homeless, then she wants her friends to contribute as well.

**Related literature**  Our work is motivated by the rise of political polarization in the United States. Boxell et al. (2017) employed nine measures of polarization and found that polarization has been rising especially among older Americans. We refer the reader to Gentzkow (2016) for an overview of polarization in 2016.

Segregation is closely related to polarization. Levy and Razin (2017) model parents’ choice between public and private schools, and they investigate how initial beliefs contribute to the long-run polarization of beliefs about the quality of public schools. Baccara and Yariv (2013) study a foundation of homophily, in which agents form groups before making a contribution to a public project. They find that stable groups must be sufficiently homogeneous. Goyal et al. (2017) experimentally test segregation through a multiple-player coordination game, and they find that subjects coordinate with those in the same designated group, even though this strategy decreases social welfare. Sethi and Yildiz (2016) study a model of social learning where agents start from heterogeneous priors and acquire qualities of different precision. Agents could observe the posterior belief of a target individual, and they find that in the long run each agent restricts her attention to
a small set of experts. The above papers study segregation, but they haven’t considered strategic communication, which we believe is crucial for the applications to religion and politics.

Several papers in political economy also consider strategic communication. For example in the model of Morris (2001) an adviser may not report the truth to a politician out of the fear of political correctness. Banerjee and Somanathan (2001) study the incentives of group members reporting information to their leader; some members might pretend they did not receive any signal. Their models have only one decision maker (e.g. a politician), whereas in our model every agent has to take an action. We model strategic communication with Bayesian persuasion. The seminal paper in this area is Kamenica and Gentzkow (2011). We consider heterogeneous priors, which relates to the work of Alonso and Camara (2016). Agents in our model also observe private signals, and persuasion of privately informed agents are studied in Kolotilin et al. (2017) and Guo and Shmaya (2017).

Our model has a matching stage before agents exchange information and take actions. Since agents care about the actions of other agents, we have a problem of matching with externality (Sasakia and Toda (1996); Pycia and Yenmez (2016)). This problem is typically challenging because an agent’s preference not only involve her own partner, but also who the other agents are matched to. However in the binary signal and binary action case, we could explicitly solve for the stable matching. A key ingredient in our proof is to interpret persuasion as a utility transfer from the receiver to the sender, so our model also relates to matching with transfers (e.g. Shapley and Shubik (1972)).

Other related papers include learning and disagreement such as Acemoglu et al. (2006), Rabin and Schrag (1999) and Fryer et al. (2015). Many papers explored non-Bayesian learning and political attitudes: Levy and Razin (2015); Ortoleva (2012); Benabou and Tirole (2006, 2011). Although non-Bayesian models help explain polarization, in this paper we focus on a rational explanation for seeking confirmatory source. We admit that non-Bayesian models could be more realistic in many contexts, but it is still worthwhile to write down a parsimonious rational model.

2 The model

Primitives  There is a state of the world \( \theta \in \Theta \). The state space is \( \Theta = \{0, 1\} \). The signal space is \( S = \{l, r\} \), where \( P[l|0] = P[r|1] = q \) for some constant \( q > \frac{1}{2} \). Thus signal \( l \) is informative about state 0, and signal \( r \) is informative about state 1.
Society has $2K$ agents, where $K$ is even. Each agent takes action $a_i \in \{0, 1\}$. Agent $i$'s utility is an additive quadratic loss function:

$$u_i(a_1, \ldots, a_{2K}; \theta) = \sum_{j=1}^{2K} (\theta - a_j)^2.$$ 

Notice that an agent wants all other agents to take an action that matches the state of the world. Examples include religious proselytising, recycling, and charity donation.

Agents have the same utility function, but they start from different priors. Priors are common knowledge. Let $p_i[\theta]$ denote agent $i$'s prior. For simplicity we use $p_i$ to represent $p_i[\theta = 1]$ when there is no confusion. Assume that half of the agents are left-leaning, and half are right-leaning:

$$p_1 \leq p_2 \leq \cdots \leq p_K < \frac{1}{2} < p_{K+1} \leq \cdots \leq p_{2K}.$$ 

Moreover we assume that no agent is obstinate:

$$\frac{(1-q)^2}{q^2 + (1-q)^2} < p_1 < p_{2K} < \frac{q^2}{q^2 + (1-q)^2}.$$ 

These bounds ensure that every agent has the potential to be persuaded. In particular if an agent sees two counter-signals (i.e., her prior is below $\frac{1}{2}$, but she sees two $r$'s), then she updates her belief to the other side of $\frac{1}{2}$.

**Matching** We consider one-to-one matchings, where each agent gets exactly one partner. Formally let $I = \{1, 2, \ldots, 2K\}$ denote the set of agents. A **one-to-one matching** is a function $\mu : I \to I$ that satisfies three conditions: 1) $\mu(i) \neq i$; 2) $\mu(\mu(i)) = i$, and 3) $\mu(i) \neq \mu(j)$ if $i \neq j$. Recall that agents are differentiated by their priors, so the matching only depends on the agents’ priors.

Each matching $\mu$ generates a communication game, and payoffs from this communication game determine an agent’s preference in the matching stage. Let $v_i(\mu)$ denote agent $i$’s expected payoff in the communication game when the matching is given by the matching $\mu$. Define agent $i$’s **preference over matchings**, denoted by $\succ_i$ as follows: we have $\mu \succ_i \mu'$ if $v_i(\mu) > v_i(\mu')$.

Since an agent’s utility depends on the entire matching $\mu$, we have a matching problem with externality, and we need to define stable matching inductively. If there are only two agents, then the matching is stable. Suppose there are more than two agents. Let
\(s(I\{i,j\})\) denote the set of stable matchings among the \(2K - 2\) agents (without \(i\) and \(j\)). Let \(\phi(i,j)\) denote the set of matchings such that \(i\) and \(j\) are matched, and the rest of the agents are matched according to a matching in \(s(I\{i,j\})\).

For a matching \(\mu\) a pair of agent \(i\) and \(j\) form a blocking pair if both the following conditions hold: 1) \(\mu(i) \neq j\), and 2) there exists a \(\mu' \in \phi(i,j)\) such that \(\mu' \succeq_i \mu\) and \(\mu' \succeq_j \mu\), and at least one inequality is strict. In words, a blocking pair means agents \(i\) and \(j\) are not matched to each other, but they both prefer a matching where they are matched to each other, and the rest form a stable matching of \(2K - 2\) agents.

A matching \(\mu\) is stable if no blocking pairs exist. One way to generate a stable matching is the deferred acceptance process. At any point an agent \(i\) (matched or unmatched) can propose to another agent \(j\). If \(j\) accepts \(i\), then \(i\) and \(j\) are temporarily matched, and they sever any previous matches. Continue this process until no one makes any further proposal. If the process stops, then the matching is stable; otherwise agents who form a blocking pair would propose to each other.

In general it is not clear whether stable matching exists or whether the deferred acceptance process will end. For our model stable matching always exists.

**Communication game** Let \(\mu(i)\) denote \(i\)'s partner from the matching stage. Each agent first gets a private i.i.d. signal \(s_i\). Then agents \(i\) and \(\mu(i)\) report to each other the signals they observe, and in this stage the communication is strategic. Finally each agent takes action based on her prior, her private signal, and her partner’s report.

We summarize the timeline for each agent:

1. Agent \(i\) commits to a reporting strategy \(\sigma_i : \{l,r\} \rightarrow \Delta\{l,r\}\).
2. Agent \(i\) gets a private signal \(s_i\) and reports \(\tilde{s}_i = \sigma_i(s_i)\) to her partner \(\mu(i)\).
3. Agent \(i\) takes action \(a_i(p_i, s_i, \tilde{s}_{\mu(i)})\), which is based on her prior, her private signal, and her partner’s report.

We briefly comment on the commitment assumption. One explanation is that the partners are engaging in long terms discussions about politics, and overtime an agent has a good idea of how honest her partner is. For example the agents engage in a continuous-time fictitious play of this communication game, and signals are verifiable after each round. Then they could figure out the empirical distribution of each other’s past reporting strategy. Another explanation is that we are primarily interested in the matching stage, and we simply choose the most tractable model for the communication game, even though
it may not be completely realistic. Finally the cheap talk model actually strengthens our result on the stability of assortative matching, as we discuss in Section 5.4.

In a subgame-perfect equilibrium each agent chooses the optimal action based on her information set:

\[
a_i(p_i, s_i, \tilde{s}_{\mu(i)}) = \arg\max_{a_i} \mathbb{E} \theta(p_i, s_i, \tilde{s}_{\mu(i)}) - (\theta - a_i)^2.
\]  

(2.1)

Agent \(i\) chooses her information disclosure rule \(\sigma_i\) to maximize her ex-ante utility:

\[
\max_{\sigma_i} \mathbb{E} \theta(s_i, s_{\mu(i)}, \tilde{s}_{\mu(i)} | p_i) \sum_{j=1}^{2K} - (\theta - a_j(p_j, s_j, \tilde{s}_{\mu(j)}))^2.
\]

Notice that agent \(i\) could only influence the action of her partner \(\mu(i)\), even though her utility depends on all agents’ actions.

**Assortative matching**  We say that a matching \(\mu\) is an **assortative matching** if for all \(i\) we have either \(p_i, p_{\mu(i)} < \frac{1}{2}\) or \(p_i, p_{\mu(i)} > \frac{1}{2}\). In other worlds the left-leaning agents are matched with other left-leaning agents, and right-leaning agents to other right-leaning agents.

We will show that every assortative matching is stable, but not all stable matchings are assortative. However if we impose two selection criteria for the set of stable matchings, then we only have assortative matchings left. The two criteria are as follows.

- A stable matching \(\mu\) **informationally dominates** another stable matching \(\mu'\) if every agent acquires at least as much information in \(\mu\) as in \(\mu'\), and at least one agent gets strictly more information in \(\mu\); moreover each agent’s utility in \(\mu\) is at least as large as her utility in \(\mu'\). A stable matching \(\mu\) is a **informationally efficient stable matching** if no other stable matching informationally dominates \(\mu\).

- A stable matching is **locally robust** if the matching remains stable after we perturb the priors by a small amount. Formally let \(s(p_1, p_2, \ldots, p_{2K})\) denote the set of stable matchings. A matching \(\mu\) is a **locally robust stable matching** if there exists a \(\delta > 0\) such that for all \(|\epsilon_i| < \delta\) we have \(\mu \in s(p_1 + \epsilon_1, p_2 + \epsilon_2, \ldots, p_{2K} + \epsilon_{2K})\).

The first criterion assumes that if an agent is indifferent between two matchings, then she prefers the one in which she acquires more information from her partner. This assumption is reasonable because after the agent plays this game, she might use her information in the future for other purposes. We use the first criterion only in one instance: we rule out matchings in which partners exchange no information at all.
For the second criterion we rule out matchings in which stability is highly sensitive to the initial priors. We focus on stable matchings which are robust to some small perturbation of the agent’s priors. This criterion is reasonable because in reality it is difficult to precisely measure an agent’s prior.

The next theorem formalizes the idea that people choose to learn from like-minded people.

**Theorem 2.1.** A stable matching that is informationally efficient and locally robust must be an assortative matching, and vice versa.

**Remark.** We could avoid the two refinement criteria, if we alternatively assume that an agent puts more weight on her own action than on other agents’ actions. In particular if agent $i$’s utility is $-(\theta - a_i)^2 - \lambda \cdot \sum_{j \neq i} (\theta - a_j)^2$ for some $\lambda < 1$, then all stable matching must be an assortative matching. We address this case in Section 5.2.

For the baseline model we assume that $\lambda = 1$ because we want to minimize the differences among the agents. We want the agents to have identical preference, to remove any suspicion that assortative matching is a built-in artifact of the model. Assuming $\lambda < 1$ strengthens the result for assortative matching. In particular if $\lambda$ is close to 0, then an agent gains negligible utility from persuading her partner, so no agent has the incentive to persuade the other side.

3 Preliminary analysis

Recall from the introduction that our main argument has two steps. First agents on the same side of 1/2 honestly exchange information with each other. Second an agent is willing to talk to the other side only if she has a stronger prior than her partner does. In this section we formally prove the first step and illustrate the second step through an example.

To see our argument more precisely, note that agent $i$’s utility is equal to $-(\theta - a_i)^2 - \sum_{j \neq i} (\theta - a_j)^2$. For the first term agent $i$ would like to seek higher quality signals from someone on the same side, and for the second term agent $i$ gets utility from persuading someone on the other side. Agent $i$ seeks to talk to the other side only if the gain from persuasion offsets the loss from signal quality for her own action. The agents who engage in such crosscutting are those who already have a strong prior – they are unlikely to be persuaded by the other side.
3.1 Incentive compatibility

We formalize the first step of our argument: like-minded agents honestly exchange information with each other. We say that agent \( i \) **tells the truth** to her partner \( \mu(i) \) if for all realizations of \( s_i \) we have \( \sigma_i(s_i) = s_i \). We say that agents \( i \) and \( \mu(i) \) are like-minded if for all realizations of \( s_i \) and \( s_{\mu(i)} \) we have \( a_i(p_i, s_i, s_{\mu(i)}) = a_{\mu(i)}(p_{\mu(i)}, s_{\mu(i)}, s_i) \). In other words like-minded agents take the same action if they honestly exchange private signals.

**Lemma 3.1.** Agent \( i \) tells the truth to \( \mu(i) \) if they are like-minded.

**Proof of sufficiency.** Agent \( i \)'s ex-ante utility (from \( \mu(i) \)'s action) is equal to

\[
\mathbb{E}_{s_i, s_{\mu(i)} \mid p_i} \mathbb{E}_{\theta \mid p_i, s_i, s_{\mu(i)}} \mathbb{E}_{\tilde{s}_i \mid s_i} [-(\theta - a_{\mu(i)}(p_{\mu(i)}, s_{\mu(i)}, \tilde{s}_i))^2].
\]

If agents \( i \) and \( \mu(i) \) are like-minded, then

\[
\mathbb{E}_{s_i, s_{\mu(i)} \mid p_i} \mathbb{E}_{\theta \mid s_i, s_{\mu(i)} \mid p_i} \mathbb{E}_{\tilde{s}_i \mid s_i} [-(\theta - a_{\mu(i)}(p_{\mu(i)}, s_{\mu(i)}, \tilde{s}_i))^2] \\
\leq \mathbb{E}_{s_i, s_{\mu(i)} \mid p_i} \mathbb{E}_{\theta \mid s_i, s_{\mu(i)} \mid p_i} [-(\theta - a_i(p_i, s_i, s_{\mu(i)}))^2] \\
= \mathbb{E}_{s_i, s_{\mu(i)} \mid p_i} \mathbb{E}_{\theta \mid s_i, s_{\mu(i)} \mid p_i} [-(\theta - a_{\mu(i)}(p_{\mu(i)}, s_{\mu(i)}, s_i))^2].
\]

The second line inequality follows from (2.1): agent \( i \)'s utility is maximized if she could choose the action for \( \mu(i) \). The third line equality follows from the definition of like-minded agents. Agent \( \mu(i) \) would take exactly the action that agent \( i \) would. The third line specifies agent \( i \)'s utility when \( \tilde{s}_i = s_i \) for all \( s_i \). Therefore agent \( i \) should tell the truth to agent \( \mu(i) \).

Lemma 3.1 can be generalized to arbitrary state space, arbitrary signal space, arbitrary action space, and arbitrary network structure. In general an agent discloses full information to like-minded neighbors – neighbors who take the actions as she would if she were in their positions. This intuition is robust.

We next characterize like-minded agents in terms of their priors. From (2.1) we deduce that \( a_i(p_i, s_i, s_{\mu(i)}) = 1 \) if and only if \( p_i[\theta = 1 \mid s_i, s_{\mu(i)}] \geq \frac{1}{2} \). Agent \( i \) takes action 1 whenever her posterior belief for state 1 is at least \( \frac{1}{2} \). Similarly we have \( a_{\mu(i)}(p_{\mu(i)}, s_{\mu(i)}, s_i) = 1 \) if and only if \( p_{\mu(i)}[\theta = 1 \mid s_{\mu(i)}, s_i] \geq \frac{1}{2} \). As a result agents \( i \) and \( \mu(i) \) are like-minded if their posteriors, after observing \( s_i \) and \( s_{\mu(i)} \), lie on the same side of \( \frac{1}{2} \). In terms of the priors we have either \( p_i, p_{\mu(i)} \in (\frac{(1-q)^2}{q^2+(1-q)^2}, \frac{1}{2}) \) or \( p_i, p_{\mu(i)} \in (\frac{1}{2}, \frac{q^2}{q^2+(1-q)^2}) \), as Figure 1 depicts. In our model we restrict the priors to the interval \( (\frac{(1-q)^2}{q^2+(1-q)^2}, \frac{q^2}{q^2+(1-q)^2}) \). Hence Lemma 3.1
implies that if agents $i$ and $\mu(i)$ have priors on the same side of $\frac{1}{2}$, then they truthfully reveal their private signals to each other.

\[
\begin{array}{c|c|c}
P_i, P_{\mu(i)} & \frac{(1-q)^2}{q^2 + (1-q)^2} & \frac{q^2}{q^2 + (1-q)^2} \\
\end{array}
\]

Figure 1: like-minded agents

**Corollary 3.2.** Agent $i$ tells the truth to $\mu(i)$ if their priors lie on the same side of $\frac{1}{2}$.

The converse of Corollary 3.2 is also true. Whenever $p_i < \frac{1}{2} < p_{\mu(i)}$, the agents benefit from persuading each other. We return to this point when we solve for the persuasion strategy in Lemma 4.1.

### 3.2 An illustrative example

We now address the second step of our argument: an agent prefers talking to the other side only if she has a stronger prior than her partner. The incentive to form a cross-link is that an agent could persuade her partner more than the other way around, but this incentive cannot hold for both sides, so the only stable matching is assortative matching. We illustrate this argument through an example with four agents.

Society has four agents: two D’s and two R’s. The D’s have prior less than $\frac{1}{2}$, and the R’s prior greater than $\frac{1}{2}$. Figure 2 presents the two possible types of matching. We claim that segmentation is the only stable matching.

![Diagram](image)

Figure 2: Segmentation is the only stable matching.

Suppose the R’s have a stronger prior than the D’s. In particular suppose $p_D = 47\%$ and $p_R = 55\%$, so R’s prior is farther from $\frac{1}{2}$. Suppose the signal quality is $P[l|0] =$
\( \Pr[r|1] = 54\% \). We pick these numbers so that one counter-signal (one \( r \)) would bring D’s belief above 50\%, whereas agent R needs two counter-signals (two \( l \)’s) to bring her belief below 50\%.

In a segmented society agents honestly exchange their signals (Corollary 3.2).

We claim that in crosscut society agents persuade each other as follows. Agent R always announces \( \tilde{s}_R = r \). Agent D announces \( l \) if \( s_D = l \), but if \( s_D = r \), then D announces \( l \) with probability 60\% and \( r \) with probability 40\%. These strategies ensure that whenever an agent observes a counter-signal, her partner’s message will bring her belief to the other side of \( \frac{1}{2} \). We omit the derivation here and defer interested readers to the proof of the general case. The main take-away is that agent R discloses no information, while agent D discloses some information.

We next compare the agents’ actions in segmented and crosscut societies. Table 1 lists the probability that an agent takes the correct action in each state. For example in state \( \theta = 0 \) agent D is 24.84\% more likely to take the correct action in a segmented society. Since agent D has a higher prior for state 0, she sees this change to the crosscut society as a loss. However agent R sees D’s change as a gain. Indeed agent R has a higher prior for state 1, and in a crosscut society D is more likely to take the correct action in state 1. Hence we could interpret a crosscut society as agent D paying agent R a utility transfer of 24.84\%.

| \( \Pr[a_D = \theta|\theta = 0] \) | segmented | crosscut | difference |
|---------------------------------|-----------|----------|------------|
| 78.84\%                         | 54\%      | -24.84\% |
| 29.16\%                         | 54\%      | +24.84\% |
| 29.16\%                         | 44.064\%  | +14.904\%|
| 78.84\%                         | 63.936\%  | -14.904\%|

Table 1: comparison

We interpret matching as follows. Every time a D is matched to an R, D pays R a transfer of 24.84\%, and R pays D a transfer of 14.904\%. The crucial observation is that R makes a smaller transfer. Think of this transfer as a measure of how susceptible the agent is to persuasion. Since R’s prior is farther away from \( \frac{1}{2} \), R is less susceptible to persuasion than D is. Indeed from the persuasion strategies we see that R discloses less information than D does, so R is more likely to persuade D than the other way around.

The R’s want to be matched with the D’s, but the D’s would reject these offers. The two D’s would propose to each other, which forces the two R’s form a link with each other, thus creating a segmented society. A crosscut society is unstable because the two
D’s form a blocking pair.

Here is the main take-away from this example. We can interpret crosscutting as matching with transfers. An agent with a stronger prior makes a smaller transfer than her partner does, so she gains from crosscutting. However this incentive can only occur in one direction; the agent with a more moderate prior would reject the cross-link, which makes segmentation the only stable matching.

4 Assortative matching: baseline model

In this section we first walk through the main ingredients of the proof of Theorem 2.1. Along the way we show that information disclosure is non-monotonic in the distance between the agents’ priors. Agents disclose more information to extreme agents because they are harder to persuade. This result stands in contrast with the cheap-talk model of Crawford and Sobel (1982), where a sender discloses coarser information as his bias grows larger.

We also consider the social planner’s problem. We propose four possible objectives the social planner might have, and we argue that assortative matching is generally inefficient for social welfare and information aggregation.

4.1 Stable matchings

Communication strategy Corollary 3.2 shows that if partners’ priors lie on the same side of $\frac{1}{2}$, they honestly exchange information with each other. We now solve for the communication strategy when agents’ priors lie on opposite sides of $\frac{1}{2}$.

Define an agent’s susceptibility to persuasion as follows.

$$f(p_i) = \begin{cases} 
\min\{1, \frac{p_i[\theta=1|s_i=r]-(1-q)}{q-p_i[\theta=1|s_i=r]}\} & \text{if } p_i < \frac{1}{2} \\
\min\{1, \frac{q-p_i[\theta=1|s_i=l]}{p_i[\theta=1|s_i=l]-(1-q)}\} & \text{if } p_i > \frac{1}{2}.
\end{cases}$$

This function specifies how easy it is to persuade agent $i$. We have $f(p)$ increases as $p$ gets closer to $\frac{1}{2}$. If an agent’s prior is close to $\frac{1}{2}$, she is highly susceptible to persuasion, whereas an extreme agent is unlikely to be persuaded. It is easy to check that $f\left(\frac{(1-q)^2}{q^2+(1-q)^2}\right) = 0$; $f(p)$ increases in the interval $\left(\frac{(1-q)^2}{q^2+(1-q)^2}, \frac{1}{2}\right)$; $f(p)$ decreases in the interval $\left(\frac{1}{2}, \frac{q^2}{q^2+(1-q)^2}\right)$, and $f\left(\frac{q^2}{q^2+(1-q)^2}\right) = 0$. 

14
The next lemma states an agent’s strategy in terms of her partner’s susceptibility to persuasion.

**Lemma 4.1.** If $p_i < \frac{1}{2} < p_{\mu(i)}$, then

- $\sigma_i(l) = l$ with probability 1, and $\sigma_i(r) = l$ with probability $\min\{f(p_{\mu(i)}), 1\}$;

- $\sigma_{\mu(i)}(r) = r$ with probability 1, and $\sigma_{\mu(i)}(l) = r$ with probability $\min\{f(p_i), 1\}$.

Notice that Lemma 4.1 is consistent with our example in Section 3.2. In the example we have $q = 54\%$ and $p_R[\theta = 1|s_R = l] = 51\%$. We obtain that $f(p_R) = \frac{54\%-51\%}{51\%-46\%} = \frac{3}{5} = 60\%$, so agent R’s susceptibility to persuasion is 60\%. On the other hand we have $p_D[\theta = 1|s_D = r] = 51\%$, and $\frac{51\%-46\%}{54\%-51\%} = \frac{5}{3} > 1$, so $f(p_D) = 1$. Hence R disclose no information, while D lies 60\% of the time when $s_D = r$.

Lemma 4.1, coupled with Corollary 3.2, imply that the amount of information disclosure is non-monotonic in the distance between the agents’ priors. In particular suppose $p_i < \frac{1}{2}$. Then $i$ discloses full information both when $p_{\mu(i)} \in (p_i, \frac{1}{2})$, no information when $p_{\mu(i)} \in (\frac{1}{2}, q)$, and an increasing amount of information as $p_{\mu(i)}$ goes from $q$ to $\frac{q^2}{q^2 + (1-q)^2}$. Agent $i$ discloses more information when her partner’s prior is either close to her or far from hers, but for different reasons. When the priors are close, the agents’ incentives are aligned, so agent $i$ discloses more information. When the priors are far apart, particularly when $p_{\mu(i)}$ gets close to $\frac{q^2}{q^2 + (1-q)^2}$, agent $\mu(i)$ becomes less susceptible to persuasion, so agent $i$ needs to disclose more information.

The above observation stands in contrast with model of Crawford and Sobel (1982). In the cheap talk model if the agents’ incentives are sufficiently far apart, then babbling is the only equilibrium (at least with the quadratic loss utility). In our model when the agents’ priors are far apart, they actually disclose more information because it is harder to persuade a more extreme agent.

**Matching** If an agent is matched to a partner on her own side (e.g., $p_i, p_{\mu(i)} < \frac{1}{2}$), she gets an untainted signal from her partner, which helps her to better infer the state of the world. On the other hand if an agent is matched to a partner on the other side ($p_i < \frac{1}{2} < p_{\mu(i)}$), then she might benefit from persuasion. An agent prefers crosscut matches if the benefit from persuasion exceeds the cost to her signal quality.

To analyze this trade-off between signal quality and benefit from persuasion, we begin with the agents’ actions in Table 2. For convenience we use $L = 0$ and $R = 1$ to denote the actions.
Note that same-side match creates correlated actions \( a_i = a_{\mu(i)} \), while actions in a crosscut match are consistent with the set of signals except in the case of \( s_i = r \) and \( s_{\mu(i)} = l \). This observation is important for our later analysis of information aggregation.

We now compare the probability an agent takes the correct action in each state. Table 3 lists the actions when \( p_i < \frac{1}{2} \), and Table 4 lists the actions when \( p_i > \frac{1}{2} \).

Blue and red parts indicate the change from a same-side matching to a crosscut matching. In particular if both agents are left-leaning \( (p_i < \frac{1}{2}, p_{\mu(i)} < \frac{1}{2}) \), then they perform well in state \( L \), but not in state \( R \). Similarly if both agents are right-leaning \( (p_i > \frac{1}{2}, p_{\mu(i)} > \frac{1}{2}) \), then they perform well in state \( R \), but not in state \( L \). These changes are analogous to the ones in Table 1. Notice that all the changes are proportional to \( q(1 - q) \). Just like in the example in Section 3.2 we could interpret a crosscut match as agent \( i \) paying her partner a transfer proportional to \( f(p_i) \), and agent \( \mu(i) \) paying a transfer proportional to \( f(p_{\mu(i)}) \). We essentially have a zero-sum game: one side gets \( f(p_{\mu(i)}) - f(p_i) \), and the other side gets \( f(p_i) - f(p_{\mu(i)}) \). This crosscut match cannot make both agents' utility positive.

We now state the agents' payoffs in a non-assortative matching. Normalize each agent’s payoff from assortative matching to 0. Tables 3 and 4 imply that each agent’s utility in

<table>
<thead>
<tr>
<th>( s_i )</th>
<th>( s_{\mu(i)} )</th>
<th>( a_i )</th>
<th>( a_{\mu(i)} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>r</td>
<td>r</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>l</td>
<td>l</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>r</td>
<td>l</td>
<td>L</td>
<td>L</td>
</tr>
<tr>
<td>r</td>
<td>l</td>
<td>f(p_i) R+(1−f(p_i)) L</td>
<td>(1−f(p_{\mu(i)})) R + f(p_{\mu(i)}) L</td>
</tr>
</tbody>
</table>

Table 2: left: \( p_i, p_{\mu(i)} < \frac{1}{2} \), right: \( p_i < \frac{1}{2} < p_{\mu(i)} \)

| \( p_i \) | \( p_{\mu(i)} \) |

Table 3: \( p_i < \frac{1}{2} \)

| \( p_i \) | \( p_{\mu(i)} \) |

Table 4: \( p_i > \frac{1}{2} \)
another matching is proportional to \( q(1 - q) \), so we can drop this term when analyzing whether an agent’s payoff is positive or negative. Suppose a matching \( \mu \) has cross-match pair \( p_j < \frac{1}{2} < p_{\mu(j)} \). For each cross-match pair \((j, \mu(j))\), agent \( i \)'s utility changes by the following amount: \( f(p_{\mu(j)}) - f(p_j) \) if \( p_i < \frac{1}{2} \), and by \( f(p_j) - f(p_{\mu(j)}) \) if \( p_i > \frac{1}{2} \). Therefore if \( p_i < \frac{1}{2} \), agent \( i \)'s utility for matching \( \mu \) is proportional to
\[
\sum_{j: p_j < \frac{1}{2} < p_{\mu(j)}} f(p_{\mu(j)}) - f(p_j).
\]
Similarly if \( p_i > \frac{1}{2} \), agent \( i \)'s utility for matching \( \mu \) is proportional to
\[
\sum_{j: p_j < \frac{1}{2} < p_{\mu(j)}} f(p_j) - f(p_{\mu(j)}).
\]
Since (4.1) and (4.2) sum up to 0, whenever there is a crosscut match, one side has a non-positive utility. If one side has a negative utility, then this side strictly prefers assortative matching. The only way for this matching to be stable is that both (4.1) and (4.2) are exactly equal to 0. This happens in two cases: either the partners have priors in the range \([1 - q, q]\), or the priors are symmetric enough. In the first case the partners exchange no information, which is ruled out by the informationally efficient criterion. In the second case if we simply perturb the priors by \( \epsilon \), then the matching would no longer be stable, and this case is ruled out by the locally robust criterion.

4.2 Efficiency

Four criteria of efficiency There is no standard definition of efficiency when agents have heterogeneous beliefs. Some recent progress on this topic include Brunnermeier et al. (2014) and Gilboa et al. (2014). Their work does not immediately apply to our setting; unlike the allocation of goods, in our setting the social planner cannot control the persuasion. However we could apply some of their insights to define efficient matching. We propose four possible objectives the social planner might have

1. Social welfare is equal to the sum of the agents’ ex-ante utility, with respect to their own priors:
\[
SW = \sum_i \mathbb{E}_{\{a_1, \ldots, a_{2K}\}, \theta | p_i} u_i(a_1, \ldots, a_{2K}; \theta).
\]

2. Social welfare is the expected sum of the agents’ utility, with respect to the social
planner’s prior \( p_{SP} \):

\[
SW = E_{\{a_1, \ldots, a_{2K}\}, \theta | p_{SP}} \sum_i u_i(a_1, \ldots, a_{2K}; \theta).
\]

Since agents have the same utility function, we could alternatively write the social welfare as

\[
SW = E_{\{a_1, \ldots, a_{2K}\}, \theta | p_{SP}} \sum_i -(\theta - a_i)^2.
\]

3. A matching \( \mu' \) dominates \( \mu \) if for all \( p_{SP} \) the social welfare defined by Criterion 2 above is (weakly) larger under \( \mu' \) than under \( \mu \), and there exists a \( p_{SP} \) for which the welfare under \( \mu' \) is strictly larger than the welfare under \( \mu \).

4. Social planner wants to maximize the probability that the majority of the agents are taking the correct action:

\[
P \left[ \sum_i -(\theta - a_i)^2 > -K \right] + \frac{1}{2} \cdot P \left[ \sum_i -(\theta - a_i)^2 = -K \right].
\]

The social planner has a neutral prior (\( p_{SP}[0] = p_{SP}[1] = \frac{1}{2} \)). She maximizes the probability that the outcome of a majority voting (with random tie-breaking) is consistent with \( \theta \), assuming the agents vote sincerely.

The first criterion is concerned with agents’ subjective utility. It is most appropriate for issues like religion; Campante and Yanagizawa-Drott (2015) find that religion increases people’s subjective happiness, while decreases economic output. The second criterion applies to cases where the social planner has a stance on a certain issue (e.g. climate change). The third criterion assumes the the social planner uses a prior-free approach. The four criterion is about information aggregation: the social planner wants to obtain the correct outcome through majority voting. This criterion is most relevant for large scale elections.

**Three other matchings** Define a left-crosscutting match \( \mu^*_L \) as follows. Let \( i_L \) denote the largest even \( i \) such that \( f(p_{i_L}) < f(p_{K+i_L}) \). Then \( \mu^*_L(1) = K + 1, \mu^*_L(2) = K + 2, \ldots, \mu^*_L(i_L) = K + i_L \), and all other agents are matched with a partner of adjacent prior. In other words \( \mu^*_L \) has as many cross links as possible so that agents on the left are still less susceptible to persuasion than agents on the right.
Define a right-crosscutting match $\mu_R^*$ as follows. Let $i_R$ denote the smallest even $i$ such that $f(p_{i_R}) > f(p_{K+i_R})$. Then $\mu_R^*(i_R) = K+i_R, \mu_R^*(i_R+1) = K+i_R+1, \ldots, \mu_R^*(K) = 2K$, and all other agents are matched with a partner of adjacent prior. In other words $\mu_R^*$ creates as many cross links as possible so that agents on the right are less susceptible than agents on the left.

Define a symmetric matching $\mu^*$ as follows. We have $\mu^*(1) = 2K, \mu^*(2) = 2K-1, \ldots, \mu^*(K) = K+1$. In other words extreme agents are matched with extreme agents, and central agents are matched with central agents, as Figure 5 depicts.

Is assortative matching efficient? We focus on a society where the priors are somewhat spread out. In particular suppose that $p_1 < 1 - q$ and $p_{2K} > q$. Moreover assume that neither $\mu_L^*$ nor $\mu_R^*$ is an assortative matching. These assumptions ensure that society has some agents with strong priors, and that crosscutting is non-degenerate. We say that such a society is somewhat diverse, for the lack of a better term.
Theorem 4.2. Suppose a society is somewhat diverse. Under all four criteria, either assortative matching is inefficient, or all matchings are efficient.

1. Ex-ante sum: all matchings are efficient if society is perfectly symmetric; i.e.,
\[ \sum_{p_i < \frac{1}{2}} \left( \frac{1}{2} - p_i \right) = \sum_{p_i > \frac{1}{2}} (p_i - \frac{1}{2}). \] Otherwise the efficient matching is either \( \mu_L^* \) or \( \mu_R^* \).

2. Expected sum under the social planner’s prior: all matchings are efficient if social planner’s prior is exactly \( \frac{1}{2} \). Otherwise the efficient matching is either \( \mu_L^* \) or \( \mu_R^* \).

3. Prior-free dominance: all matchings are efficient.

4. Majority voting: the social planner prefers \( \mu^* \) to assortative matching.

From the perspective of information aggregation (Criterion 4), assortative matching is inefficient. This result is not surprising because it is well-known in the social learning literature that homophily slows down the diffusion of information (Golub and Jackson (2012)). In our setting assortative matching creates correlated errors – if an agent takes the wrong action, then her partner also takes the wrong action, as we see in Table 2. On the other hand a matching crosscut pairs have actions that are more consistent with the set of signals. In particular Table 2 shows that in a crosscutting match we have \( a_i = s_i \) except when the signals are \( (r, l) \), so agents are almost voting informatively. As a result information aggregation is more efficient in matching \( \mu^* \).

The inefficiency of assortative matching under Criterion 4 seems counter-intuitive. Indeed in an assortative matching each agent acquires two i.i.d. signals, so individuals are better informed than they would be in any other matching. How could a society with more informed agents fails to efficiently aggregate information? The problem boils down to correlated actions: more informed agents are taking correlated actions in an assortative matching. A related paper by Levy and Razin (2015) studies the opposite effect: less informed voters could aggregate information more efficiently because they ignore the correlation between their signals.

For the other three criteria we see that assortative matching is sometimes efficient and sometimes inefficient. For Criterion 1 when an extreme agent is matched with a moderate agent on the other side, the utility gain from the extreme agent exceeds the utility loss from the moderate agent. Thus matchings \( \mu_L^* \) and \( \mu_R^* \) both matches as many extreme agents to moderate agents as possible. The only case when assortative matching becomes efficient is that the priors are symmetric, in which case the social welfare is the same for all matchings.
For Criterion 2 if the social planner’s prior is greater than $\frac{1}{2}$, then she wants the extreme agents on the right to persuade the moderate agents on the left, so $\mu^*_R$ is efficient. Similarly if her prior is less than $\frac{1}{2}$, then she wants extreme agents on the left to persuade moderate agents on the right, so $\mu^*_L$ is efficient. If the social planner has a prior exactly at $\frac{1}{2}$, then she is indifferent between all matchings.

For Criterion 3 assortative matching is efficient, but prior-free dominance gives an extremely weak notion of efficiency. In fact every matching is efficient under this criterion. Indeed if matching $\mu$ dominates $\mu'$ when the social planner’s prior is less than $\frac{1}{2}$, then we have $\mu'$ dominates $\mu$ when the social planner’s prior is greater than $\frac{1}{2}$. No matching could dominate another one for all possible priors.

In summary assortative matching is inefficient for information aggregation, and for social welfare it is inefficient except in cases where all matchings are equally efficient. Therefore we conclude that when efficiency is a meaningful property, assortative matching is inefficient. One potential solution for restoring efficiency is to allow agents to make transfers during the match (Shapley and Shubik (1972)), but in reality it seems unrealistic for people to pay someone for a political discussion. Hence the main take-away from this section is that assortative matching is stable, but socially inefficient.

5 Assortative matching: extended models

This section serves as a robustness check to Theorem 2.1. Our argument for assortative matching continues to hold after we modify the baseline model in four different ways. First there are multiple states instead of binary states. Second an agent puts more weights on her own action than on other agents’ actions. Third we consider heterogeneous preferences instead of heterogeneous priors. Finally we use cheap talk instead of Bayesian persuasion to model information exchange.

5.1 Multiple states

Suppose state space is a discrete subset of the interval $[0,1]$. Signals and actions are still binary. Signals satisfy the Monotone Likelihood Ratio Property: $\frac{P(r|\theta)}{P(r'|\theta)}$ is increasing in $\theta$. 
Define agent $i$’s prior expectations for $\theta$ as follows:

$$
e_i = \sum_{\theta} \mathbb{P}[lr|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_i[\theta]$$

$$e_i^+ = \sum_{\theta} \mathbb{P}[rr|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_i[\theta] > 0$$

$$e_i^- = \sum_{\theta} \mathbb{P}[ll|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_i[\theta] < 0$$

Assume all agents are non-obstinate, so that $e_i^+ > 0$ and $e_i^- < 0$ for all $i$. Instead of ranking the agents by their priors $p_i$ as we did in the binary case, we now rank the agents by their prior expectations $e_i$. Assume exactly $K$ agents have negative values for $e_i$, and the other $K$ agents have positive values for $e_i$.

We say a matching $\mu$ is an assortative matching if we have $e_i \cdot e_{\mu(i)} > 0$ for all $i$. We claim that Theorem 2.1 continues to hold under this new definition of assortative matching.

First we show that if $e_i \cdot e_{\mu(i)} > 0$, then agents $i$ and $\mu(i)$ honestly exchange their signals. Notice that an agent takes action 0 if her expectation of $\theta$ is less than $\frac{1}{2}$. In other words after she observes signal $s$ she takes action 0 if $\sum_{\theta} \mathbb{P}[\theta|s] \cdot p_i[\theta] < \frac{1}{2}$. This condition is equivalent to $\sum_{\theta} \mathbb{P}[s|\theta] \cdot p_i[\theta] \cdot (\theta - \frac{1}{2}) < 0$. If $e_i \cdot e_{\mu(i)} > 0$, then these two agents would take the same action after they see each other’s signals. In particular they both will take action 0 unless $s_i = s_{\mu(i)} = r$. By the same argument as Lemma 3.1 we deduce that agents $i$ and $\mu(i)$ honestly exchange information with each other.

We now consider the persuasion strategy when $e_i < 0 < e_{\mu(i)}$. Agent $i$ would announce $\tilde{s}_i = l$ whenever $s_i = l$, and she would announce $\tilde{s}_i = l$ with some probability $\sigma_i$ if $s_i = r$. Agent $i$ needs to pick $\sigma_i$ such that her partner prefers action 0 to action 1, which means $p_{\mu(i)}[s_i = l|\tilde{s}_i = l] \geq \frac{e_{\mu(i)}}{e_{\mu(i)} - e_{\mu(i)}}$. We know that $p_{\mu(i)}[s_i = l|\tilde{s}_i = l] = \frac{p_{\mu(i)}[l]}{p_{\mu(i)}[l] + \sigma_i(1 - p_{\mu(i)}[l])}$, which implies that

$$
\sigma_i \leq \frac{p_{\mu(i)}[l]}{p_{\mu(i)}[r]} \cdot \frac{e_{\mu(i)}^-}{e_{\mu(i)}^+} = \frac{\sum_{\theta} \mathbb{P}[l|\theta] \cdot p_{\mu(i)}[\theta]}{\sum_{\theta} \mathbb{P}[r|\theta] \cdot p_{\mu(i)}[\theta]} \cdot \frac{-\sum_{\theta} \mathbb{P}[l|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_{\mu(i)}[\theta]}{-\sum_{\theta} \mathbb{P}[r|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_{\mu(i)}[\theta]} = f(p_{\mu(i)}).
$$

The optimal choice for $\sigma_i$ is the minimum of $f(p_{\mu(i)})$ and 1. Similarly we define

$$f(p_i) = \frac{p_i[r]}{p_i[l]} \cdot \frac{e_i^+}{e_i^-} = \frac{\sum_{\theta} \mathbb{P}[r|\theta] \cdot p_i[\theta]}{\sum_{\theta} \mathbb{P}[l|\theta] \cdot p_i[\theta]} \cdot \frac{-\sum_{\theta} \mathbb{P}[r|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_i[\theta]}{-\sum_{\theta} \mathbb{P}[l|\theta] \cdot (\theta - \frac{1}{2}) \cdot p_i[\theta]}.$$
Agent \( \mu(i) \) announces \( l \) with probability 1 if \( s_{\mu(i)} = l \), and announces \( l \) with probability \( \min\{1, f(p_i)\} \) if \( s_{\mu(i)} = r \).

Now consider the matching stage. Normalize everyone’s payoff in an assortative matching to 0. Suppose \( \mu \) is not an assortative matching, and \( e_i < 0 < e_{\mu(i)} \) for some agent \( i \). Then the actions of \( i \) and \( \mu(i) \) would change when the signals are \( s_i = r \) and \( s_{\mu(i)} = l \). Agent \( i \)'s utility from persuading \( \mu(i) \) is proportional to \( \sum_{\theta} p_i[\theta|lr] \cdot (\theta - \frac{1}{2}) \cdot f(p_{\mu(i)}) \), and there is also a probability \( f(p_i) \) that agent \( i \) will be persuaded by her partner. Hence agent \( i \)'s utility in matching \( \mu \) from this crosscutting pair is proportional to

\[
(f(p_{\mu(i)}) - f(p_i)) \cdot \sum_{\theta} p_i[\theta|lr] \cdot (\theta - \frac{1}{2}).
\]

Similarly agent \( \mu(i) \)'s utility from this cross-match is proportional to

\[
(f(p_i) - f(p_{\mu(i)})) \cdot \sum_{\theta} p_{\mu(i)}[\theta|lr] \cdot (\theta - \frac{1}{2}).
\]

The above two terms (agent \( i \)'s and \( \mu(i) \)'s utility) cannot both be positive. If \( f(p_{\mu(i)}) - f(p_i) > 0 \), then agent \( \mu(i) \) is more susceptible to persuasion, and agent \( i \) prefers this crossmatch, but agent \( \mu(i) \) would reject. Such a matching cannot be stable. The argument for assortative matching follows exactly the same logic as Theorem 2.1.

### 5.2 More weight on agent’s own action

We return to the baseline model in Section 2, but we modify the utility function as follows:

\[
u_i(a_1, \ldots, a_{2K}; \theta) = -(\theta - a_i) - \lambda \cdot \sum_{j \neq i} (\theta - a_j)^2
\]

for some \( \lambda \in (0, 1) \). If \( \lambda \) is close to 0, then each agent primarily cares about her own action, whereas if \( \lambda \) is close to 1, then each agent puts significant weights on other agents’ actions. Our baseline model corresponds to the limit case of \( \lambda \to 1 \).

We claim that this setting strengthens our result for assortative matching. In particular Theorem 2.1 holds even without the refinement criteria. When \( \lambda \in (0, 1) \), every stable matching must be an assortative matching.

Notice that the communication strategies are exactly the same as before. Indeed agent \( i \)'s persuasion problem is to maximize the expected value of \( -\lambda \cdot (\theta - a_{\mu(i)}) \), and the coefficient \( \lambda \) does not change the persuasion strategy. Therefore we could continue the
analysis directly from Tables 3 and 4.

Normalize each agent’s payoff from assortative matching to 0. Tables 3 and 4 imply that each agent’s utility in another matching is proportional to $q(1 - q)$, so we can drop this term when analyzing whether an agent’s payoff is positive or negative. Suppose a matching $\mu$ has cross-match pair $p_j < \frac{1}{2} < p_{\mu(j)}$. For each cross-match pair $(j, \mu(j))$, agent $i$’s utility changes by the following amount: $f(p_{\mu(j)}) - f(p_j)$ (or $f(p_{\mu(j)}) - \frac{1}{2} f(p_j)$ if $i = j$) if $p_i < \frac{1}{2}$, and by $f(p_j) - f(p_{\mu(j)})$ (or $f(p_j) - \frac{1}{2} f(p_{\mu(j)})$ if $i = \mu(j)$) if $p_i > \frac{1}{2}$. Therefore if $p_i < \frac{1}{2}$, agent $i$’s utility for matching $\mu$ is proportional to

$$\sum_{j: p_j < \frac{1}{2} < p_{\mu(j)}} f(p_{\mu(j)}) - f(p_j) - 1_{i=j} \cdot (1/\lambda - 1) f(p_j).$$

(5.1)

Similarly if $p_i > \frac{1}{2}$, agent $i$’s utility for matching $\mu$ is proportional to

$$\sum_{j: p_j > \frac{1}{2} < p_{\mu(j)}} f(p_j) - f(p_{\mu(j)}) - 1_{i=j} \cdot (1/\lambda - 1) f(p_{\mu(j)}).$$

(5.2)

Since $\lambda < 1$, the sum of (5.1) and (5.2) is negative if $p_i < \frac{1}{2} < p_{\mu(i)}$. Notice that the difference between this case and the baseline model is that we cannot have both (5.1) and (5.2) equal to 0, which spares us from using the refinement criteria. Whenever there is a cross-match, at least one side has a negative utility. In particular, since $K$ is even, any non-assortative matching $\mu$ must have at least two pairs of cross-matches, denoted by $(j_1, \mu(j_1))$ and $(j_2, \mu(j_2))$. Without loss of generality assume that $\sum_{j: p_j < \frac{1}{2} < p_{\mu(j)}} f(p_{\mu(j)}) - f(p_j) \leq 0$. Then we have $v_{j_1} < 0$ and $v_{j_2} < 0$. Agents $j_1$ and $j_2$ form a blocking pair. We thus obtain assortative matching without any refinement criteria.

One special case of this model is that when $\lambda$ is close to 0, each agent primarily cares about her own action. Thus each agent is mainly concerned with seeking high quality signals instead of persuading someone on the other side. To see this point more precisely, as $\lambda \to 0$ we have (5.1) and (5.2) both go to negative infinity. As a result nobody wants to talk to the other side.

5.3 Heterogeneous preferences

So far we have considered a model where agents are distinguished by their priors about the state of the world. In many applications agents also have different preferences (i.e., conflicts of interests over a certain policy). In this section we apply our model to the
setting where agents start from a common prior, but heterogeneous preferences. We derive an analogous result for assortative matching.

We modify our baseline model as follows. All agents start from a common prior \( p_i = \frac{1}{2} \). Agent \( i \) has a preference \( v_i \), which satisfies

\[
-\frac{q - \frac{1}{2}}{q^2 + (1 - q)^2} < v_1 < \cdots < v_K < 0 < v_{K+1} < \cdots < v_{2K} < \frac{q - \frac{1}{2}}{q^2 + (1 - q)^2}.
\]

A matching \( \mu \) is an \textbf{assortative matching} if \( v_i \cdot v_{\mu(i)} > 0 \) for all \( i \). Actions are still binary, and agent \( i \)'s utility is defined as

\[
u_i(a_1, \ldots, a_{2K}; \theta) = \sum_j -(a_j - \theta + v_i)^2.
\]

Think of \( v_i \) as agent \( i \)'s bias. Agent \( i \) wants everyone to take an action that matches the state of the world with respect to her own bias.

Agent \( i \) takes action 1 if her posterior belief exceeds \( v_i + \frac{1}{2} \). In other words we have \( a_i(v_i, s_i, \hat{s}_{\mu(i)}) = 1 \) whenever \( P[\theta = 1|s_i, \hat{s}_{\mu(i)}] \geq v_i + \frac{1}{2} \). The derivation is straightforward. Let \( p \) denote the agent’s posterior belief for state 1. Her expected payoff from action 0 is equal to \( \mathbb{E}[-(0 - \theta + 1)^2] = -p(-1 + v_i)^2 - (1 - p)v_i^2 = p(2v_i - 1) - v_i^2 \). Her expected payoff from action 1 is equal to \( \mathbb{E}[-(1 - \theta + v_i)^2] = -pv_i^2 - (1 - p)(1 + v_i)^2 = p(2v_i + 1) - (1 + v_i)^2 \). The agent takes action 1 if \( p(2v_i + 1) - (1 + v_i)^2 \geq p(2v_i - 1) - v_i^2 \), which is equivalent to \( p \geq v_i + \frac{1}{2} \).

The persuasion strategy is analogous to the case of heterogeneous priors. Define the susceptibility to persuasion as follows:

\[
f(v) = \frac{q^2}{q^2 + (1 - q)^2} - \left(\frac{|v|}{2}\right) - \frac{q - \frac{1}{2}}{q^2 + (1 - q)^2 - q}.
\]

The function \( f \) is non-decreasing as \( v \) gets closer to 0.

\textbf{Lemma 5.1. We have two cases for persuasion:}

- \textbf{Truth-telling:} if \( v_i \cdot v_{\mu(i)} > 0 \), then \( \sigma_i(s_i) \equiv s_i \).

- \textbf{Distortion:} if \( v_i < 0 < v_{\mu(i)} \), then \( \sigma_i(r) = r \) with probability 1, and \( \sigma_i(l) = r \) with probability \( \min\{1, f(v_{\mu(i)})\} \).

The argument for assortative matching follows the same logic as in the baseline model. For an agent \( i \) with \( v_i < 0 \), when changing from an assortative matching to a partner on
the other side \((v_{\mu(i)} > 0)\), agent \(i\)'s change in action occurs when the signals are \((r,l)\), in which case her action changes from \(f(v_i) \cdot L\) to \(f(v_i) \cdot R\). The associated change in utility is listed in Table 5.

<table>
<thead>
<tr>
<th>(\theta = 0)</th>
<th>assortative</th>
<th>crosscutting</th>
</tr>
</thead>
<tbody>
<tr>
<td>(-f(v_i) \cdot v_i^2)</td>
<td>(-f(v_i) \cdot (1 + v_i)^2)</td>
<td></td>
</tr>
<tr>
<td>(-f(v_i) \cdot (-1 + v_i)^2)</td>
<td>(-f(v_i) \cdot v_i^2)</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: utility change for an agent with \(v_i < 0\)

Agent \(i\)'s loss in utility for each cross-link is \(q(1 - q) \cdot f(v_i) \cdot [(1 + v_i)^2 - (-1 + v_i)^2]\), which is equal to \(q(1 - q) \cdot f(v_i) \cdot 4v_i\). Dropping the coefficients we deduce that for each cross-link agent \(i\)'s loss in utility is proportional to \(f(v_i) \cdot v_i\). Similarly her gain in utility (from \(\mu(i)\)'s change in action) is proportional to \(f(v_{\mu(i)}) \cdot v_i\). Hence if \(v_i < 0\), then for any matching \(\mu\) agent \(i\)'s utility is proportional to

\[
v_i \cdot \sum_{j: v_j < 0 < v_{\mu(j)}} f(v_{\mu(j)}) - f(v_j),
\]

and if \(v_i > 0\) then her utility is proportional to

\[
v_i \cdot \sum_{j: v_j < 0 < v_{\mu(j)}} f(v_j) - f(v_{\mu(j)}).
\]

Note that for any matching \(\mu\) with cross-links, if agents on the left side of zero get positive utility, then agents on the right side of zero get negative utility. This zero-sum utility implies that \(\mu\) cannot be stable because at least agents who get negative utility prefer assortative matching, and if they get matched to the other side, they form a blocking pair. More precisely if \(j_1 < 0 < \mu(j_1)\) and \(j_2 < 0 < \mu(j_2)\), then we cannot have all four agents having positive utility. Suppose \(j_1\) and \(j_2\) have negative utility. Then they form a blocking pair. Hence any stable matching that is informationally efficient and locally robust must be an assortative matching.

### 5.4 Cheap talk

In the baseline model we use Bayesian persuasion to model the information exchange. A common criticism of Bayesian persuasion is the commitment assumption: in our case agent \(i\) commits to a reporting strategy \(\sigma_i\) before the information exchange game. In this
section we remove the commitment assumption. Instead we assume that the agents play an equilibrium strategy, as modeled in Crawford and Sobel (1982).

The cheap talk model strengthens our result on assortative matching. Without the commitment assumption an agent gets less utility from persuading her partner, because her strategy set now shrinks significantly. Removing commitment reduces an agent’s incentive to talk to someone on the other side, which strengthens the stability of assortative matching.

In the cheap model whenever \( p_i < \frac{1}{2} < p_{\mu(i)} \) the only communication equilibrium is babbling. Indeed in a non-babbling equilibrium an agent must be indifferent between reporting \( l \) or \( r \), but then the agent’s report won’t change her partner’s action, which means her partner ignores the message. Hence babbling is the only equilibrium.

Normalize everyone’s payoff in an assortative matching to 0. We claim that every cross-match will make all agents’ utility negative. Suppose \( p_i < \frac{1}{2} < p_{\mu(i)} \). If \( p_i \in (1 - q, \frac{1}{2}) \), then agent \( i \)’s change in action is the same as Table 2. In particular when the signals are \( s_i = r \) and \( s_{\mu(i)} = l \), agent \( i \) takes action 0 instead of action 1 as she would in an assortative matching, and she sees this change as a loss. If \( p_i < 1 - q \), then agent \( i \) will always take action 0 in a babbling equilibrium, which is worse than an assortative matching because she prefers take action 1 when \( s_i = s_{\mu(i)} = r \). We next examine \( i \)’s benefit from persuading \( \mu(i) \). If \( p_{\mu(i)} > q \), then in a babbling equilibrium \( \mu(i) \) will always take action 1. Agent \( i \) gets nothing from persuasion; in fact she loses utility as well because agent \( \mu(i) \) takes action 1 even if \( s_i = s_{\mu(i)} = l \), in which case she would have taken action 0 in an assortative matching. As a result any cross-match that involves agents outside the interval \((1 - q, q)\) creates negative externality for all agents. The only cross-match could involve agents in the interval \((1 - q, q)\), just like in the model with Bayesian persuasion. We could rule out this type of matching in two ways: using the informationally efficient criterion: an agent is indifferent between this matching and assortative matching, but she acquires strictly more information in an assortative matching. Therefore any informationally efficient stable matching must be an assortative matching.

In our setting the cheap talk model has two notable differences with the Bayesian persuasion model. First in the Bayesian persuasion model if \( \mu(i) \) has a strong prior (i.e., far from \( \frac{1}{2} \)), then agent \( i \) would disclose more information to \( \mu(i) \) because persuading an extreme guy is difficult. In the cheap talk model however agent \( i \) would simply babble to \( \mu(i) \). Though we have only considered binary signals, in the cheap talk model, at least the uniform quadratic case, as the sender and the receiver’s incentives grow farther apart, babbling becomes the only equilibrium. The intuition from cheap talk is that agent \( i \)
gives less information as \( \mu(i) \) gets more extreme because their incentives are mis-aligned, whereas in Bayesian persuasion agent \( i \) would release more information because it is harder to persuade \( \mu(i) \).

The second difference is that in the Bayesian persuasion model agents with strong priors prefer to talk to the other side, whereas in the cheap talk model these agents prefer to talk to their own side. The key difference is that in cheap talk model when the priors are far apart agents disclose little information, so there is little benefit to persuasion. In contrast in the Bayesian persuasion model agents with a strong prior would benefit greatly from persuasion. In reality the partisans are more likely to seek information from the a wide spectrum of sources (Gentzkow and Shapiro (2011)), so the Bayesian persuasion model seems more intuitive.

6 Communication on a network

So far we have focused on one-to-one matchings, where each agent has only one neighbor to communicate with. In reality people have multiple friends to discuss religion and political issues. In this section we allow the agents to have multiple neighbors and investigate their communication strategies on a network.

A major challenge for extending our results from the one-to-one matching is that on a network an agent’s persuasion strategy depends on the equilibrium strategies of other agents. Indeed on a network we have a communication game with multiple senders, so each agent’s strategy depends on other agents’ messages. Therefore we are no longer studying a bilateral communication game; we need to characterize the equilibrium strategy in a game with multiple senders, which is a difficult task in general.

In this section we report two progress toward this goal. First we extend Lemma 3.1 to more general settings, noting that “like-minded people tell the truth” is a robust result. Second we show in some special cases that “stable network” involves homophily.

6.1 Network communication model

**States, actions, utility** A society has \( K \) individuals. They each has a prior about the state of the world \( \theta \in \Theta \). Let \( p_i \in \Delta(\Theta) \) denote agent \( i \)'s prior, which is common knowledge. Let \( A \) denote the action space, and let \( a_i \in A \) denote the action of agent \( i \). Agent \( i \)'s utility depends on the state of the world as well as everyone's action: \( u_i(a_1, \ldots, a_K; \theta) \). We put no restriction on this utility function. Here are some examples:
• Jury voting: \( \Theta = A = \{0, 1\} \), and

\[
u_i(a_1, \ldots, a_K; \theta) = \begin{cases} 
1 & \text{if } a_1 = \cdots = a_K = \theta = 0 \\
1 & \text{if } \theta = 1 \text{ and } a_i = 1 \text{ for some } i \\
0 & \text{otherwise}
\end{cases}
\]

• Majority voting: \( \Theta = A = \{0, 1\} \), and

\[
u_i(a_1, \ldots, a_K; \theta) = \begin{cases} 
1 & \text{if } \sum_i - (\theta - a_i)^2 > -\frac{K}{2} \\
0 & \text{otherwise}
\end{cases}
\]

• Coordination (e.g. Golub and Morris (2017); Hagenbach and Koessler (2010)):

\[
u_i(a_1, \ldots, a_K; \theta) = -(\theta - a_i)^2 + \sum_{j \neq i} -\lambda_{ij}(a_i - a_j)^2.
\]

• Matching the state:

\[
u_i(a_1, \ldots, a_K; \theta) = -(\theta - a_i)^2 + \sum_{j \neq i} -\lambda_{ij}(\theta - a_j)^2.
\]

So far we have been working with this utility function for the one-to-one matching.

**Communication** Society is an undirected graph. Let \( N(i) \) denote the neighbors of agent \( i \). Agent \( i \) can only communicate with her neighbors.

Let \( S \) denote the signal space. Assume all signals are i.i.d. Each agent gets one private signal; let \( s_i \) denote agent \( i \)'s signal.

Let \( M \) denote the message space. After the agents observe their private signal, they send a message (privately) to each of their neighbors. Since communication is private, an agent could send different messages to different neighbors. More precisely agent \( i \) reports \( \tilde{s}_{ij} \in \Delta M \) to neighbor \( j \in N(i) \), where \( \tilde{s}_{ij} \) is a function of \( s_i \).

Agent \( i \) receives a message \( \tilde{s}_{ji} \) from each neighbor \( j \in N(i) \), where \( \tilde{s}_{ji} \) is a function of \( s_j \). Agent \( i \)'s posterior depends on her prior \( p_i \), her private signal \( s_i \), and her neighbor’s messages \( \{\tilde{s}_{ji}\}_{j \in N(i)} \). In particular her posterior for state \( \theta \) is \( p_i[\theta|s_i; \{\tilde{s}_{ji}\}_{j \in N(i)}] \). She takes an action \( a_i \) based on this posterior.

In summary the timeline is as follows:
1. Agent $i$ commits to a reporting strategy $\sigma_{ij} : S \rightarrow \Delta(M)$ for each $j \in N(i)$.

2. Agent $i$ gets a private signal $s_i$ and reports $\tilde{s}_{ij} = \sigma_{ij}(s_i)$ to $j$.

3. Agent $i$ takes action $a_i(p_i, s_i, \{\tilde{s}_{ji}\}_{j \in N(i)})$, where $\tilde{s}_{ji}$ is $j$’s report to $i$.

**Subgame perfect equilibrium** First each agent must take the optimal action given her information set. Let $S_i = \{s_i, \{\tilde{s}_{ji}\}_{j \in N(i)}\}$ denote the set of agent $i$’s private signal and her neighbor’s messages. Let $a_{-i}$ denote the actions of other agents. Then agent $i$ takes the following action:

$$a_i(p_i, S_i) = \arg\max_{a_i} \mathbb{E}_{\theta, a_{-i} | p_i, S_i} [u_i(a_i, a_{-i}; \theta)].$$

Moreover agent $i$ chooses her reporting strategy $\sigma_{ij}$ to maximize her ex-ante utility:

$$\max_{\{\sigma_{ij}\}_{j \in N(i)}} \mathbb{E}_{\theta, s_i, \{s_j\}_{j \in N(i)}} [u_i(a_i(p_i, S_i), a_{-i}(p_{-i}, S_{-i}); \theta)]. \quad (6.1)$$

Assume that when agent $i$ prefers honesty for tie-breaking: if she is indifferent between an honest message ($\tilde{s}_{ij} = s_i$) and another message, she prefers the honest one.

**Remark.** At first glance agent $i$ could only influence the action of her neighbors. In particular agent $i$’s message $\tilde{s}_{ij}$ enters into $j$’s information set $S_j$ and thereby affects $j$’s action $a_j(p_j, S_j)$. In fact agent $i$ could also influence the actions of non-neighbors, because their strategies depend on the actions of $i$’s neighbors.

### 6.2 A general result on incentive compatibility

We say that agent $i$ tells the truth to agent $j$ if $\sigma_{ij}(s_i) = s_i$ for all $s_i$. In other words agent $i$ honestly reports her private signal to agent $j$. If all neighbors tell the truth to $j$, then $j$ updates her belief based on $|N(j)| + 1$ i.i.d. signals.

We identify a sufficient condition for agent $i$ to tell the truth to agent $j$. This sufficient condition basically says $i$ and $j$ think alike. Specifically let $S$ denote agent $j$’s information set. We define

$$a_{ij}(p, S) = \arg\max_{a_j} \mathbb{E}_{\theta, a_{-j} | p, S} [u_i(a_j, a_{-j}; \theta)].$$

This $a_{ij}(p, S)$ represents the action agent $i$ would take if she were in agent $j$’s position. The sufficient condition says that agent $i$ would take exactly the same action as $j$ takes under any circumstance.
More precisely we fix the strategies \( \{ \sigma_{kj} \}_{k \in N(j) \setminus i} \) of \( j \)'s other neighbors. Suppose that for all \( S_j = \{ s_j, \{ \tilde{s}_{kj} \}_{k \in N(j) \setminus i} \} \) we have \( a_j(p_j, S_j, s_i) = a_{ij}(p_i, S_j, s_i) \). Then agent \( i \) is a like-minded neighbor under strategies \( \{ \sigma_{kj} \}_{k \in N(j) \setminus i} \). This condition is sufficient for agent \( i \) to tell the truth to agent \( j \).

**Theorem 6.1.** Fix the strategies \( \{ \sigma_{kj} \}_{k \in N(j) \setminus i} \) of agent \( j \)'s other neighbors. If agent \( i \) is a like-minded neighbor under strategies \( \{ \sigma_{kj} \}_{k \in N(j) \setminus i} \), then agent \( i \) tells the truth to \( j \).

Theorem 6.1 generalizes Lemma 3.1. Note that we put no assumption on the primitives. This result holds for general utility function and arbitrary network structure. The proof only depends on the fact that agents \( i \) and \( j \) have their incentive aligned. Hence the intuition that “like-minded people tell the truth” is a robust result.

**Proof.** The ex-ante utility of \( i \) is equal to

\[
\mathbb{E}_{s_i, S_j | p_i} \mathbb{E}_{\theta, a_{-j} | p_i, s_i, S_j} \left( \mathbb{E}_{\tilde{s}_{ij} | s_i} \left[ u_i(a_j(p_j, S_j, \tilde{s}_{ij}), a_{-j}; \theta) \right] \right).
\]

By definition of \( a_{ij} \) the optimal \( a_j \) in the parenthesis should be \( a_{ij}(p_i, S_j, s_i) \). We get

\[
\begin{align*}
\mathbb{E}_{s_i, S_j | p_i} \mathbb{E}_{\theta, a_{-j} | p_i, s_i, S_j} \left( \mathbb{E}_{\tilde{s}_{ij} | s_i} \left[ u_i(a_j(p_j, S_j, \tilde{s}_{ij}), a_{-j}; \theta) \right] \right) \\
\leq \mathbb{E}_{s_i, S_j | p_i} \mathbb{E}_{\theta, a_{-j} | p_i, s_i, S_j} \left[ u_i(a_{ij}(p_i, S_j, s_i), a_{-j}; \theta) \right] \\
= \mathbb{E}_{s_i, S_j | p_i} \mathbb{E}_{\theta, a_{-j} | p_i, s_i, S_j} \left[ u_i(a_j(p_j, S_j, s_i), a_{-j}; \theta) \right]
\end{align*}
\]

Therefore the optimal reporting strategy is to set \( \tilde{s}_{ij} = s_i \). \( \square \)

### 6.3 Stable networks and homophily

In our model agents always prefer to have an additional link in the network. In the worst case they could ignore the messages from this link and send no message through the link. It is not clear how we should define stable networks, so we focus on a simple case.

Suppose \( K \) is even. Each agent has exactly \( \frac{K}{2} - 1 \) neighbors. There are two types of priors: \( p_L \) and \( p_R \). Half of the agents have prior \( p_L \), and the other half \( p_R \). In an **assortative network**, agents with the same prior are connected among themselves, so we have homophily. We say agents \( i \) and \( j \) form a **blocking pair** if \( i \) is connected to \( i' \); \( j \) is connected to \( j' \); neither \( i \) and \( j \) nor \( i' \) and \( j' \) are connected, but both \( i \) and \( j \) prefer the links \( (i, j), (i', j') \) over \( (i, i'), (j, j') \). A network is **stable** if there does not exist any blocking pairs. In this section we demonstrate in two cases that assortative networks are stable.
6.3.1 Lexicographic preference

Suppose each agent primarily cares about their own action, such as sincere voting in juries. Agent $i$’s utility can be separated into $u_i(a_i, \theta)$ and $u_i(a_{-i}, \theta)$. Assume that $u_i(a_i, \theta) = |a_i - \theta|$ is equal to the Euclidean distance between agent $i$’s action and the true state. Assume that $u_i(a_{-i}, \theta) = \sum_{j \neq i} |a_j - \theta|$ is equal to the sum of Euclidean distances between other agents’ actions and the true state. Agent $i$ has lexicographic preference over these two components:

$$u_i(a_1, \ldots, a_K; \theta) = \text{LEX}\{u_i(a_i, \theta), u_i(a_{-i}, \theta)\}.$$  

Agent $i$ first maximizes $u_i(a_i, \theta)$, and conditional on $u_i(a_i, \theta)$ agent $i$ maximizes $u_i(a_{-i}, \theta)$. Note that the first component determines her incentive to seek high quality information, while the second component gives her the incentive to persuade other agents.

**Proposition 6.2.** If all agents have lexicographic preference, then assortative network is stable.

Proposition 6.2 follows directly from Theorem 6.1. Indeed whenever agents $i$ and $j$ have the same prior, they honestly exchange information with each other because they both want the other to take an action close to the true state. Hence in an assortative network each agent acquires $|K/2|$ i.i.d. signals, which is the largest amount of information they could obtain in any network. Since an agent first wants to maximize the distance between her own action and the true state, she wants to acquire $|K/2|$ i.i.d. signals. The only reason she would talk to the other side is that she could benefit from persuasion. However agents on the other side would refuse to incur any loss in signal quality, so they wouldn’t want to be matched with her.

Note that lexicographic preference simplifies our analysis for stability. We could avoid the trade-off between signal quality and persuasion: under lexicographic preference an agent first and foremost wants to maximize the quality of her information. Whenever an agent could benefit from persuasion, her target would reject this link. With more general utility functions, an agent would have to evaluate the trade-off between signal quality and benefit to persuasion, which would require us to solve for the communication strategies. We solve a special case of this problem in the next section.

6.3.2 Binary states, signals, actions

Suppose the states, signals, and actions are all binary as in the baseline model of one-to-one matching. Moreover assume that agents have the common utility function $\sum_i -(a_i -$
We show that our argument for Theorem 2.1 carries over to the analysis of stable networks, when there are only two types of priors \( p_L \) and \( p_R \) where \( p_L < \frac{1}{2} < p_R \).

We first solve for the communication strategies. From Theorem 6.1 we know that if all of \( i \)'s neighbors have the same prior as \( i \), then they honestly report their signals. Hence in an assortative network each agent acquires \( |K/2| \) i.i.d. signals.

We next show that all other matchings involve some distortion of signals. We assume that \( p_L \) and \( p_R \) are not “obstinate”. Let \( S_i \) denote a collection of \( |N(i)| \) i.i.d. signals of \( r \). Let \( S_l \) denote a collection of \( |N(i)| \) i.i.d. signals of \( l \). We say that agent \( i \) is obstinate if either \( p_l[\theta = 1|S_l] < 1 - q \) or \( p_l[\theta = 1|S_l] > q \). An obstinate agent would take action 0 even if she sees \( |N(i)| \) occurrences of \( r \).

**Lemma 6.3.** Suppose no agent is obstinate. All of agent \( i \)'s neighbors honestly report their signals if and only if they all have the same prior as agent \( i \).

We next analyze the agents’ persuasion strategies when agents with different priors form a link. Let \( S_i = \{s_i, \{\tilde{s}_{k_i}\}_{k \in N(i) \setminus j}\} \) denote agent \( i \)'s information set. It consists of agent \( i \)'s private signal \( s_i \) as well as the messages from \( i \)'s other neighbors. Agent \( j \) benefits from persuasion if and only if there exists a realization of \( S_i \) such that \( p_j[\theta = 1|s_j = r, S_i] \in (0, \frac{1}{2}) \) and \( p_l[\theta = 1|s_j = r, S_i] \in \left(\frac{1}{2}, \frac{q^2}{q^2 + (1-q)^2}\right) \), as shown in Figure 6. Indeed, if no such \( S_i \) exists, then agent \( i \) (conditional on other neighbors’ messages) is obstinate about state 1. Now let’s suppose such \( S_i \) exists. Then agent \( j \) wants to persuade \( i \) to take action 0 whenever \( i \) has access to these information sets \( S_i \).

![Figure 6: \( j \) is not a like-minded neighbor of \( i \)](image)

Here is agent \( j \)'s best-response strategy. She finds the critical point at which she wants to persuade \( i \). Let

\[
p_{ji}^* = \max_{S_i} p_l[\theta = 1|s_i, \{\tilde{s}_{k_i}\}_{k \in N(i) \setminus j}] 
\]

such that \( p_j[\theta = 1|s_j = r, \{s_i, \{\tilde{s}_{k_i}\}_{k \in N(i) \setminus j}\}] \in (0, \frac{1}{2}) \) and \( p_l[\theta = 1|s_j = r, \{s_i, \{\tilde{s}_{k_i}\}_{k \in N(i) \setminus j}\}] \in \left(\frac{1}{2}, \frac{q^2}{q^2 + (1-q)^2}\right) \). Whenever agent \( i \)'s interim belief \( p_l[\theta = 1|\{s_i, \{\tilde{s}_{k_i}\}_{k \in N(i) \setminus j}\}] \) falls between \( \frac{1}{2} \) and \( p_{ji}^* \), agent \( j \) wants to persuade agent \( i \) to take action 0. If agent \( i \)'s interim belief is outside this interval, then agent \( j \)'s message won’t affect \( i \)'s action. From Lemma 4.1
we obtain that agent \( j \)'s optimal strategy is as follows: \( \sigma_{ji}(l) = l \) with probability 1, and \( \sigma_{ji}(r) = l \) with probability \( \min\{1, \frac{q-p_{**}}{p_{**}-(1-q)}\} \).

**Lemma 6.4.** Each agent \( i \) selects a \( p_{ij}^* \in [1-q, q] \) for each neighbor \( j \in N(i) \).

- If \( p_i[\theta = 1] < p_j[\theta = 1] \), then \( \sigma_{ij}(l) = l \) with probability 1, and \( \sigma_{ij}(r) = l \) with probability \( \min\{1, \frac{q-p_{ij}^*}{p_{ij}^*-(1-q)}\} \).
- If \( p_i[\theta = 1] > p_j[\theta = 1] \), then \( \sigma_{ij}(r) = r \) with probability 1, and \( \sigma_{ij}(l) = r \) with probability \( \min\{1, \frac{p_{ij}^*-(1-q)}{q-p_{ij}^*}\} \).

Lemma 6.4 actually holds for general network with general priors, but for the analysis of stable network we focus on the special case when there are only two types of priors. Half of the agents have prior \( p_L \), and the other half have prior \( p_R \). Assortative network is stable.

**Proposition 6.5.** With binary states, signals, and actions, assortative network is stable.

Note that Proposition 6.5 gives a sufficient conditions for stability. The converse is much harder to study. For the persuasion strategy we identified the “cut-off” strategies in Lemma 6.4, but we have not derived any close-form solutions for \( p_{ij}^* \). These cut-offs depend on the network structure. On the other hand Lemma 6.4 and Theorem 6.1 both apply to general networks, and they could provide useful tool for future investigation on stable networks.

## 7 Non-assortative matching

Up to this point we have argued that assorative matching arises in a variety of contexts. We now explore the limitations of our model: when does non-assortative matching occur?

Our baseline model has several assumptions: (i) binary states, (ii) binary signals, (iii) binary actions, (iv) one-to-one matching, (v) identical weights on all agents’ actions, (vi) heterogeneous priors instead of heterogeneous preferences, and (vii) commitment assumption in Bayesian persuasion. In Sections 5 and 6 we argued that our result on assortative matching is robust to assumptions (i), (iv), (v), (vi), (vii). It remains to check assumptions (ii) and (iii).

In this section we demonstrate the limitations of the binary signals and binary actions assumptions. We also discuss the case of multi-dimensional learning. We construct simple counter-examples with four agents that give rise to non-assortative matching. All of our
counter-examples have the following feature: agents 1 and 2 have the same prior; 3 and 4 have the same prior, but all agent prefer the crosscut matching \((1, 3), (2, 4)\) over the assortative matching \((1, 2), (3, 4)\).

### 7.1 Multiple actions

We return to the model in Section 2, but assume that there are three possible actions: \(A = \{0, \frac{1}{2}, 1\}\). An agent takes action 0 if her posterior belief or state 1 is less than \(\frac{1}{4}\), action \(\frac{1}{2}\) if her belief is between \(\frac{1}{4}\) and \(\frac{3}{4}\), and action 1 if her belief is above \(\frac{3}{4}\).

Society has 4 agents. Two of them have prior \(\frac{1}{4} - \epsilon\), and the other two have prior \(\frac{3}{4} + \epsilon\): we have \(p_1 = p_2 = \frac{1}{4} - \epsilon\) and \(p_3 = p_4 = \frac{3}{4} + \epsilon\). Assume that \(q\) is close to \(\frac{1}{2}\) so that \(p_1[\theta = 1| r]\) and \(p_1[\theta = 1| rr]\) both lie in the interval \((\frac{1}{4}, \frac{3}{4})\). Hence agents 1 and 2 never take action 1, and agents 3 and 4 never take action 0.

We claim that \((1, 3), (2, 4)\) is a stable matching, while \((1, 2), (3, 4)\) is unstable. The first matching is crosscutting, and the second one is assortative.

![Figure 7: agents 1 and 2 have the same prior; 3 and 4 have the same prior.](image)

In an assortative matching, agents honestly exchange signals and take the same action (Theorem 6.1). Agents 1 and 2 both take action 0, unless \(s_1 = s_2 = r\), in which case they both take action \(\frac{1}{2}\). Agents 3 and 4 both take action 1, unless \(s_3 = s_4 = l\), in which case they both take action \(\frac{1}{2}\).

In a crosscutting matching, agents exchange no information; they babble to each other. Indeed if agent 3 observes \(s_3 = r\), she will take action 1 regardless of agent 1’s message. If agent 3 observes \(s_3 = l\), then she will take action \(\frac{1}{2}\), and even if agent 1 persuades her that \(s_1 = l\), agent 3 will still take action \(\frac{1}{2}\). Agent 1 can never persuade 3 to take action 0, and in the case of \(s_3 = l\) agent 3 would already take action \(\frac{1}{2}\). Therefore agent 1 releases no information to agent 3, and vice versa.

We claim that all agents prefer crosscutting over assortative matching. Normalize everyone’s payoff in an assortative matching to 0. The difference in actions between these two matchings occur when \(s_1 = r\) and \(s_2 = l\). Agent 1 takes action 0 in assortative matching, but takes action \(\frac{1}{2}\) in crosscutting, so her utility loss is proportional to \(\mathbb{E}_{p_1} [- (\theta - \ldots]}

35
On the other hand agent 1 gains from persuading agent 3 from taking action 1 to action \( \frac{1}{2} \). Agent 1’s benefit from persuasion is proportional to \( -\mathbb{E}_{p_1}[(\theta - \frac{1}{2})^2] - \mathbb{E}_{p_1}[(\theta - 1)^2] \). Hence agent 1’s net payoff in a crosscutting match is proportional to

\[
-\mathbb{E}_{p_1}[(\theta - 0)^2] + \mathbb{E}_{p_1}[(\theta - \frac{1}{2})^2] + \mathbb{E}_{p_1}[(\theta - \frac{1}{2})^2] - \mathbb{E}_{p_1}[(\theta - 1)^2].
\]

Since the quadratic loss function is concave, agent 1’s utility is positive. Indeed we have

\[
\mathbb{E}_p[-(\theta - a)^2] = p(1-a)^2 + (1-p)(0-a)^2 = (a-p)^2 + p(1-p),
\]

which is concave in \( a \). Consequently agent 1 strictly prefers crosscutting over assortative matching. The same argument applies symmetrically to agent 3. Hence in an assortative matching agents 1 and 3 form a blocking pair. In a crosscutting match, however, no pair of agents would form a blocking pair because they are both worse off in an assortative matching.

The key difference between binary actions and multiple actions is that in the binary-action model no agent wants to talk an agent with a strong prior. With binary actions an agent must persuade her partner from prior 0.76 down to 0.5, but with multiple actions she only needs to persuade her partner from 0.76 to 0.75. Now agents see more benefits to persuasion.

### 7.2 Multiple signals

Suppose the states and actions are both binary. Signal space is \( S = [0, 1] \). We have \( P[s|\theta = 1] = 2s \) and \( P[s|\theta = 0] = 2 - 2s \). Note that the distribution of \( S \) satisfies the Monotone Likelihood Ratio Property: \( \frac{P[s|\theta = 1]}{P[s|\theta = 0]} \) is increasing in \( s \), so higher realization of \( s \) indicates higher probability for state 1.

Society has four agents. Agents 1 and 2 have prior \( p_1 = p_2 = \epsilon \). Agents 3 and 4 have prior \( p_3 = p_4 = 1 - \epsilon \). We claim that all agents prefer the matching (1, 3), (2, 4) over the matching (1, 2), (3, 4).

Let \( s \) and \( s' \) denote two i.i.d. signals. For small enough \( \epsilon \) there exists a \( \delta < \frac{1}{2} \) such that \( p_1[\theta = 1|s, s'] \geq 1/2 \) only if both \( s, s' > 1 - \delta \), and \( p_3[\theta = 1|s, s'] \leq 1/2 \) only if both \( s, s' < \delta \). For small enough \( \epsilon \) if agents 1 and 3 persuade each other, agent 1 only needs to focus on the region \( s_1, s_3 < \delta \), and agent 3 only needs to focus on the region \( s_1, s_3 > 1 - \delta \).

Agent 1’s utility gain from persuading agent 3 is equal to

\[
\int_{s_1, s_3 < \delta} p_1[s_1, s_3] \cdot \mathbb{E}_{\theta|s_1, s_3, p_1} \mathbb{E}_{a_3(s_1, s_3)}[(\theta - a_3(s_1, s_3))^2] - (\theta - a_3(s_1, s_3))^2] \tag{7.1}
\]
Agent 1’s utility loss from being persuaded by agent 3 is equal to

$$\int_{s_1, s_3 > 1 - \delta} p_1(s_1, s_3) \cdot \mathbb{E}_{\theta|s_1, s_3, p_1} \mathbb{E}_{a_1(s_1, s_3)} [(\theta - a_1(s_1, s_3))^2 - (\theta - a_1(s_1, s_3))^2].$$  

(7.2)

We claim that for small enough $\epsilon$ agent 1’s net utility is positive, so she prefers crosscutting over assortative matching. The same argument applies to other agents by symmetry.

**Lemma 7.1.** For small enough $\epsilon$ we have (7.1) is greater than (7.2).

The intuition of this example is straightforward: each agent gets persuaded in the states that she thinks are unlikely, and she could persuade her partner in the states in which she thinks are highly likely. Agent 1 thinks small $s_1, s_3$ are unlikely, while large $s_1, s_3$ are highly likely. Continuous signals makes it easier to persuade a partner with a strong prior than with binary signals, which gives agents more incentive for crosscutting.

### 7.3 Multi-dimensional model

Our argument for assortative matching continues to hold with multiple states (see Section 5.1). However when the states are multi-dimensional, our argument no longer holds. The intuition is straightforward. In the one-dimensional case an agent is willing to talk to the other side only if she has a stronger prior than her partner, and this incentive cannot hold in both directions. However with multiple dimensions, an agent could be extreme in dimension 1 and moderate in dimension 2, while her partner is extreme in dimension 2 and moderate in dimension 1. Then both of them have the incentive to persuade each other in the dimensions for which they have a strong prior.

**Model** State of the world is two-dimensional: $\Theta = \{0, 1\} \times \{0, 1\}$. Signal space is $S = \{l, r\} \times \{l, r\}$. Signals for each dimension are independent, and $P[s_d = l|\theta_d = 0] = P[s_d = r|\theta_d = 1] = q$ for both $d = 1$ and $d = 2$. Action space is $A = \Theta$.

Each agent has a prior for each state. Let $p_{i1}$ and $p_{i2}$ denote agent $i$’s prior for each dimension. Let $a_i = (a_{i1}, a_{i2})$ denote agent $i$’s action in each dimension. Utility is the sum of quadratic loss: $u_i(a_1, \ldots, a_{2K}; \theta) = \sum_j - (\theta_1 - a_{j1})^2 - (\theta_2 - a_{j2})^2$.

**Non-assortative matching** Society has four agents. Agents 1 and 2 have identical prior: $p_{i1} = \frac{(1-q)^2}{q^2+(1-q)^2} + \epsilon$ and $p_{i2} = \frac{1}{2} + \epsilon$ for $i = 1, 2$. Agents 3 and 4 have identical prior: $p_{i1} = \frac{1}{2} + \epsilon$ and $p_{i2} = \frac{(1-q)^2}{q^2+(1-q)^2} + \epsilon$. Agents 1 and 2 have a strong prior for dimension 1.
and nearly neutral prior for dimension 2, while agents 3 and 4 have exactly the opposite prior.

In an assortative matching agents 1 and 2 talk to each other, and 3 and 4 talk to each other. We claim that every agent prefers crosscutting: agent 1 talks to 3, and 2 talks to 4.

Since the utility is separable in each dimension, and signals are independent across the two dimensions, the persuasion strategy remains the same as the single-dimensional case in Lemma 4.1. In particular when agents 1 and 3 talk to each other, agent 1 babbles in dimension 1, and in dimension 2 if agent 1 sees a signal \( r \), she lies with probability \( f(p_{32}) \). Similarly agent 3 babbles in dimension 2, and in dimension 1 if she sees a signal \( r \), she lies with probability \( f(p_{11}) \).

Normalize all agents’ payoffs in an assortative matching to 0. Then in a crosscutting matching agent 1’s utility is proportional to

\[
\left| \frac{1}{2} - \frac{(1 - q)^2}{q^2 + (1 - q)^2} - \epsilon \right| \cdot (1 - f(p_{32}) - \frac{1}{2} - (\frac{1}{2} + \epsilon)) \cdot (1 - f(p_{11})).
\]

We have \( f(p_{32}) = f(p_{11}) < 1 \). As \( \epsilon \to 0 \) the second term disappears, while the first term remains positive. Hence agent 1 prefers crosscutting to assortative matching. The same argument applies to agent 3. The key idea is that each agent loses some utility in the dimension in which her prior is nearly neutral, but she compensates for this loss through her gain from persuasion in the other dimension in which she has a strong prior.

### 7.4 Discussion

We constructed all three counter-examples using the same strategy: increase an agent’s benefit from persuasion. Every agent has to balance the trade-off between seeking truthful signals and persuading her partner. Models in Sections 5 and 6 weaken an agent’s incentive to persuade the other side, while the examples in this section seek to strengthen this incentive. Multiple actions, multiple signals, and multiple dimensions all make it easier for an agent to persuade someone on the other side, even if her partner has a strong prior.

Do these counter-examples nullify our result on assortative matching? One should not immediately dismiss the stability of assortative matching. Although the counter-examples in this section are straightforward, they mostly involve extreme agents from opposite sides persuading each other. We believe that even with multiple signals and actions, the moderate agents still prefer to talk to other moderates. In particular if an
agent’s prior is exactly $\frac{1}{2}$, then she is indifferent between the actions that other agents take, so she gains nothing from persuasion. Hence we should still expect the agents who engage in crosscutting to have strong priors.

The binary-action model is more appropriate for the U.S. politics because most elections have only two viable alternatives. A potential empirical test is to compare polarization in Europe and polarization in America. Many European countries have multiple parties, which corresponds to the multiple-action example.

The binary-signal assumption is hard to test empirically, because we have to test whether people are looking for yes/no answers or trying to understand the nuances of an certain issue. This question is more about psychology than rational learning. One potential evidence for binary signals comes from a recent study (Gabielkov et al., 2016), which finds that 59% of social media users read headlines only. However it is still difficult to test how these people interpret the headlines.

The counter-examples imply that people are more willing to seek the opinion of the opposite side if there is a continuum of beliefs. One example is the Creation vs. Evolution debate. It used to be a binary choice. During the Scope Monkey Trial in 1925 evolution was seen as a denial to the Christian faith. Several states passed laws prohibiting the teaching of evolution. Today however the theory of Intelligent Design creates many intermediate positions. Although two-thirds of Americans still want public schools to teach Creationism, the majority of these people support Intelligent design, and only a quarter of Americans support teaching Creationism only (Berkman and Plutzer (2010)).

8 Conclusion

We have written a model in which assortative matching is stable, but is inefficient for social welfare and information aggregation. We also examined the limitations of our model. We propose two questions for future research. The first is a technical question: can we extend our analysis to more general network settings? In Section 6 we have only considered a simple case of stable networks. It would be interesting to see whether in general a stable network would involve a large degree of homophily. In addition dynamic learning is also an important aspect to model, even though it’s not clear ex-ante what the best model is. Ideally we want to explain not only why the U.S. electorate is polarized, but why it gets more polarized over time.

The second question concerns an alternative model of learning based on social identity. For example an agent might avoid discussing the negative consequences of affirmative
action because she doesn’t want to be labeled as a racist. Benabou and Tirole (2011) consider a model where an agent is trying to infer what type of person she is, but their model is primarily about a single agent’s learning. If we consider interactions between agents who deeply care about their identity, could homophily arise?

References


A Proof of Theorem 2.1

We clarify that the proof of Theorem 2.1 depends on Lemma 4.1, and the proof of Lemma 4.1 is self-contained.
Proof of Theorem 2.1. Normalize each agent’s payoff from an assortative matching to 0. Since in an assortative matching partners always disclose full information, an agent’s payoff in any assortative matching is identical. Therefore this normalization is well-defined.

Let \( \mu \) be a non-assortative matching. Then agent \( i \)'s utility from matching \( \mu \) is given by (4.1) and (4.2). In particular if \( p_i < \frac{1}{2} \), then

\[
v_i(\mu) = \left| \frac{1}{2} - p_i \right| \cdot \sum_{j : p_j < \frac{1}{2} < p_{\mu(j)}} f(p_{\mu(j)}) - f(p_j),
\]

and if \( p_i' > \frac{1}{2} \), then

\[
v_i'(\mu) = \left| \frac{1}{2} - p_i' \right| \cdot \sum_{j : p_j < \frac{1}{2} < p_{\mu(j)}} f(p_j) - f(p_{\mu(j)}).
\]

We cannot have both \( v_i(\mu) > 0 \) and \( v_i'(\mu) > 0 \).

We prove the theorem by induction. We strengthen the inductive hypothesis as follows: if there is an even number of agents on each side of \( \frac{1}{2} \), then in any stable matching every agent has a payoff of 0.

Suppose for contradiction that not agents have 0 utility in matching \( \mu \). Since utility is zero-sum, agents on one side of \( \frac{1}{2} \) have positive utility, and agents on the other side of \( \frac{1}{2} \) have negative utility. Then any pair of agents on the side with negative utility form a blocking pair. Indeed, by the inductive hypothesis, the other \( 2K - 2 \) agents form a matching in which everyone has payoff 0. Since this pair is on the same side of \( \frac{1}{2} \), including this pair in the matching still results in everyone having payoff 0, and they both prefer this match over the one that gives them negative utility.

We next apply the refinement criteria. If \( \mu \) contains a pair such that \( p_i \in (1 - q, \frac{1}{2}) \) and \( p_{\mu(i)} \in (\frac{1}{2}, q) \), then agents \( i \) and \( \mu(i) \) exchange no information with each other because \( f(p_i) = f(p_{\mu(i)}) = 1 \). This type of matching is ruled out by the informationally efficient criterion. Now suppose \( \mu \) contains a crosscut pair such that \( p_i < 1 - q \) or \( p_{\mu(i)} > q \), then perturbing either \( p_i \) or \( p_{\mu(i)} \) by \( \epsilon \) would change \( f(p_i) \) or \( f(p_{\mu(i)}) \), which makes \( v_i(\mu) \) no longer equal to 0. Thus \( \mu \) fails the locally robust criterion. Therefore if a stable matching that is both informationally efficient and locally robust must be an assortative matching.

We now show that assortative matching is stable. Suppose \( p_i < \frac{1}{2} < p_j \). Then \( i \) and \( j \) cannot form a blocking pair in an assortative matching. Suppose \( i \) and \( j \) form a link with each other. Let \( \mu \) denote the new match. Regardless of what \( \mu \) is, we cannot have both
\(v_i(\mu) > 0\) and \(v_j(\mu) > 0\). At least one of the agents \(i\) and \(j\) prefer assortative matching over \(\mu\). Hence no blocking pair exists in an assortative matching. It’s easy to see that assortative matching is informationally efficient because every agent get two i.i.d. signals, and it is locally robust because perturbing the priors by \(\epsilon\) still makes \(p_i\) and \(p_{\mu(i)}\) on the same side of \(\frac{1}{2}\). \(\square\)

**B Omitted proofs in Section 4**

We begin with an auxiliary lemma about the persuasion strategy.

**Lemma B.1.** Suppose \(p_i < \frac{1}{2} < p_{\mu(i)}\). We have \(\sigma_i(l) = l\) with probability 1, and \(\sigma_i(r) = l\) with probability

\[
\min\left\{1, \frac{1}{2} - \frac{p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l]}{p_{\mu(i)}[\theta = 1|s_i = r, s_{\mu(i)} = l]} - \frac{1}{2} \cdot \frac{p_{\mu(i)}[s_i = l|s_{\mu(i)} = l]}{p_{\mu(i)}[s_i = r|s_{\mu(i)} = l]} \right\}.
\]

**Proof.** If \(s_{\mu(i)} = r\), then \(\mu(i)\) will take action 1 regardless of \(\tilde{s}_i\), so persuasion is irrelevant. If \(s_{\mu(i)} = l\), then agent \(i\) should take action 0 regardless of \(s_i\); even if \(s_i = r\), we still have \(p_i[\theta = 1|s_{\mu(i)} = l, s_i = r] < \frac{1}{2}\). Hence when \(s_{\mu(i)} = l\), agent \(i\) persuades \(\mu(i)\) to take action 0.

Suppose \(\mu(i)\) observes \(s_{\mu(i)} = l\), and agent \(i\) announces that \(\tilde{s}_i = l\). What does \(\mu(i)\) infer about \(s_i\)? We have

\[
p_{\mu(i)}[s_i = l|\tilde{s}_i = l, s_{\mu(i)} = l] = \frac{p_{\mu(i)}[s_i = l|s_{\mu(i)} = l]}{1 + p_{\mu(i)}[\tilde{s}_i = l|s_i = r] \cdot \frac{1}{2} - \frac{p_{\mu(i)}[s_i = r|s_{\mu(i)} = l]}{p_{\mu(i)}[s_i = l|s_{\mu(i)} = l]} - \frac{1}{2} p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l]}.
\]
Consequently $\mu(i)$ infers about $\theta$ as follows:

\[
p_{\mu(i)}[\theta = 1|\bar{s}_i = l, s_{\mu(i)} = l] = p_{\mu(i)}[s_i = l|\bar{s}_i = l, s_{\mu(i)} = l] \cdot p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l] + p_{\mu(i)}[s_i = r|\bar{s}_i = l, s_{\mu(i)} = l] \cdot p_{\mu(i)}[\theta = 1|s_i = r, s_{\mu(i)} = l] = 1 \cdot p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l] + \frac{1}{1 - \frac{1}{2}p_{\mu(i)}[\theta = 1|s_i = r, s_{\mu(i)} = l] - \frac{1}{2}p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l]}
\]

\[
= \frac{1}{2} - \frac{p_{\mu(i)}[\theta = 1|s_i = l, s_{\mu(i)} = l] - \frac{1}{2}p_{\mu(i)}[\theta = 1|s_i = r, s_{\mu(i)} = l]}{p_{\mu(i)}[\theta = 1|s_i = r, s_{\mu(i)} = l] + (1 - \bar{p}_{\mu(i)})q} \cdot \frac{\tilde{p}_{\mu(i)}(1 - q) + (1 - \bar{p}_{\mu(i)})q}{\tilde{p}_{\mu(i)}q + (1 - \bar{p}_{\mu(i)(1 - q)}}
\]

\[
= \frac{(1 - \tilde{p}_{\mu(i)})q - \tilde{p}_{\mu(i)}(1 - q)}{\tilde{p}_{\mu(i)}q - (1 - \bar{p}_{\mu(i)})(1 - q)}
\]

The last term, if less than 1, is exactly equal to $f(p_{\mu(i)})$, so $\sigma_i(r) = l$ with probability $f(p_{\mu(i)})$. Proof for $\mu(i)$’s strategy is symmetric. \hfill\(\square\)

**Proof of Lemma 4.1.** Let $\tilde{p}_{\mu(i)} = p_{\mu(i)}[\theta = 1|s_i = l]$. We simplify the formula from Lemma B.1.

\[
\]
and if \( p_{i'} > \frac{1}{2} \), then

\[
v_{i'}(\mu) = \left| \frac{1}{2} - p_{i'} \right| \cdot \sum_{j : p_{j} < \frac{1}{2} < p_{\mu(j)}} f(p_{j}) - f(p_{\mu(j)}).
\]

Social welfare is equal to

\[
\sum_{i} v_{i}(\mu) = \left[ \sum_{j : p_{j} < \frac{1}{2} < p_{\mu(j)}} f(p_{j}) - f(p_{\mu(j)}) \right] \cdot \left[ \sum_{p_{i} < \frac{1}{2}} (\frac{1}{2} - p_{i}) - \sum_{p_{i} > \frac{1}{2}} (p_{i} - \frac{1}{2}) \right].
\]

The second term is fixed. If it is 0, then the social welfare is 0 for any matching. If it is non-zero, then the efficient matching maximizes or minimizes the first term. The first term is maximized with \( \mu_{R}^{*} \) and minimized with \( \mu_{L}^{*} \).

For Criterion 2 the social planner maximizes \( \sum_{i} - (\theta - a_{i})^{2} \). Then formula (4.1) and (4.2) give the social planner’s utility is equal to \( (\frac{1}{2} - p_{SP}) \cdot \sum_{j : p_{j} < \frac{1}{2} < p_{\mu(j)}} f(p_{\mu(j)}) - f(p_{j}) \).

If \( p_{SP} = \frac{1}{2} \), the social planner is indifferent between all matchings. If \( p_{SP} < \frac{1}{2} \), then \( \mu_{L}^{*} \) maximizes her utility, and if \( p_{SP} > \frac{1}{2} \), then \( \mu_{R}^{*} \) maximizes her utility.

For Criterion 3 there cannot exist another matching \( \mu \) that always gives the social planner a positive utility. Indeed from the preceding discussion in Criterion 2 we see that the social planner’s utility for \( \mu \) depends on her prior. Let’s say her utility if \( x \) when \( p_{SP} < \frac{1}{2} \), then her utility becomes \(-x\) when \( p_{SP} > \frac{1}{2} \). Hence one of those utilities must be negative, and no matching could dominate an efficient matching.

We prove the last result by induction on \( K \). For the base case suppose there are four agents. If the four signals consist of two \( l \)’s and two \( r \)’s, then the conditional distribution of \( \theta \) is half-half, so the probability of having the correct outcome is \( \frac{1}{2} \) for any matching.

If the signals are four \( l \)’s or four \( r \)’s, then for any matching the agents vote for four \( L \)’s or four \( R \)’s. We are left with the cases when there are three \( l \)’s and one \( r \) or three \( r \)’s and one \( l \).

<table>
<thead>
<tr>
<th>signals</th>
<th>crosscut</th>
<th>assortative</th>
</tr>
</thead>
<tbody>
<tr>
<td>r, l, l, l</td>
<td>RRLL/LLLL/LRLL/RLLL</td>
<td>LLLL</td>
</tr>
<tr>
<td>l, l, l, r</td>
<td>LRLL</td>
<td>LLLL</td>
</tr>
<tr>
<td>l, l, r, l</td>
<td>LLRR/LLLL/LLL/RRL</td>
<td>LLRR</td>
</tr>
<tr>
<td>l, l, l, r</td>
<td>LLLL</td>
<td>LLRR</td>
</tr>
</tbody>
</table>

Table 6

one \( l \). Table 6 shows the votes for each society when there are three \( l \)’s and one \( r \). For the crosscut, the priors are \( p_{l} < \frac{1}{2}, p_{\mu} > \frac{1}{2}, p_{j} < \frac{1}{2}, \) and \( p_{\mu(j)} > \frac{1}{2} \). For assortative, the priors
are $p_i < \frac{1}{2}$, $p_{\mu(i)} < \frac{1}{2}$, $p_j > \frac{1}{2}$, and $p_{\mu(j)} > \frac{1}{2}$. The actions are tabulated based on Table 2. Conditional on having three $l$’s and one $r$, the assortative matching votes for LLLL with probability $\frac{1}{2}$ and LLRR with probability $\frac{1}{2}$, so the probability that majority voting from assortative matching gets to $R$ is $\frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$. For the crosscut society the probability that the votes are tied (2$L$’s and 2$R$’s) is less than $\frac{1}{2}$, so the probability that majority voting outputs $R$ is less than $\frac{1}{2}$. Thus conditional on having three $l$’s and one $r$, the crosscut society is more likely to vote for $L$. By the symmetric argument, conditional on having three $r$’s and one $l$, the crosscut society is more likely to vote for $R$.

We now move to the inductive step. Suppose there are $2K$ agents. Then we partition them into $K/2$ sub-societies of size four, and each society mimics the base case. If any sub-society gets 2$l$’s and 2$r$’s, then the conditional probability for $\theta$ is half-half, so no matter how the whole society votes, probability of correct outcome is always $\frac{1}{2}$. By the Principle of Inclusion and Exclusion, in cases where at least one sub-society gets 2$l$’s and 2$r$’s, then probability that majority outputs the correct outcome is $\frac{1}{2}$. We only need to focus on the case when no sub-society gets 2$l$’s and 2$r$’s. If one sub-society gets 3$l$’s and 1$r$, and another one gets 3$r$’s and 1$l$, then the conditional probability for $\theta$ is again $\frac{1}{2}$, and majority voting always output the correct state with probability $\frac{1}{2}$. We can further restrict our attention to cases where all sub-societies get $llll$, $rrrr$, or one of $3l + 1r$ or $3r + 1l$ (but not both).

Without loss of generality suppose all societies get $llll$, $rrrr$, or $3l + 1r$. For the first two sets of signals, the agents are always voting LLLL and RRRR. Only the third set generates different votes for each matching. Suppose the total number of $r$’s is more than the total number of $l$’s. Then the only way for the society to vote $L$ instead of $R$ is when some sub-societies who get $3l + 1r$ vote LLLL. Conditional on 3$l$’s and 1$r$, the probability of LLLL for assortative matching is $\frac{1}{2}$, but the probability for crosscut is less than $\frac{1}{2}$. Hence assortative matching is more likely to vote the wrong state when there are more $r$’s than $l$’s. Similarly if the total number of $l$’s is more than the total number of $r$’s, then the only way for society to vote $R$ instead of $L$ is that some societies who got $3l + r$ vote 2$L$’s and 2$R$’s. This probability is $\frac{1}{2}$ for assortative matching, but less than $\frac{1}{2}$ for crosscut. Thus conditional on more $l$’s than $r$’s, assortative matching is more likely to vote for the wrong state.

\end{proof}

C Omitted proofs in Section 5

Proof of Lemma 5.1. We have three cases.
First case: \( v_i \cdot v_{\mu(i)} > 0 \). This case corresponds to the “like-minded” agents. Agents \( i \) and \( \mu(i) \) will report to each other their true signals, because they would take the same action after observing each other’s signals. Agent \( \mu(i) \) would already take an action that maximizes agent \( i \)’s utility, so there is no need for persuasion.

Second case: \( v_i < 0 < v_{\mu(i)} \) and \( v_{\mu(i)} < q - \frac{1}{2} \). Then \( i \) babbles: \( \sigma_i(s_i) \equiv r \). Indeed, if \( s_{\mu(i)} = l \), then \( \mu(i) \)’s posterior for state \( 1 \) is less than \( q \), and \( \mu(i) \) takes action 0 because \( v_{\mu(i)} + \frac{1}{2} = q \). If \( s_{\mu(i)} = r \), then \( \mu(i) \)’s posterior for state \( 1 \) becomes \( q = v_{\mu(i)} + \frac{1}{2} \), so \( \mu(i) \) takes action 1 even without \( i \)’s message. Thus agent \( i \) simply babbles.

Third case: \( v_i < 0 \) and \( v_{\mu(i)} > q - \frac{1}{2} \). Then \( \sigma_i(r) = r \) with probability 1, and \( \sigma_i(l) = r \) with probability \( f(v) \). Here is why. If \( \mu(i) \) observes \( s_{\mu(i)} = l \), then persuasion is useless because she will take action 1 regardless of \( \tilde{s}_i \). The only case persuasion matters is when \( \mu(i) \) observes \( s_{\mu(i)} = r \). Suppose \( \mu(i) \) believes that \( \mathbb{P}[s_i = r | \tilde{s}_i = r] = \pi \). Then upon seeing \( \tilde{s}_i = r \), agent \( \mu(i) \) infers that with probability \( \pi \) we have \( s_i = r \), in which case her posterior becomes \( \mathbb{P}[\theta = 1 | s_i = r] = \frac{q^2}{q^2 + (1 - q)^2} \); and with probability \( 1 - \pi \) we have \( s_i = l \), in which case her posterior becomes \( \mathbb{P}[\theta = 1 | s_i = l] = \frac{1}{2} \). Agent \( i \) needs to choose \( \pi \) such that \( \pi \cdot \frac{q^2}{q^2 + (1 - q)^2} + (1 - \pi) \cdot \frac{1}{2} \leq \frac{1}{2} + v \). Therefore the optimal choice for \( \pi \) is to set the equality holds, which means \( \pi \) is equal to \( \frac{(v + \frac{1}{2}) - \frac{1}{2}}{q^2 + (1 - q)^2} \). Now let’s solve for \( \mathbb{P}[\tilde{s}_i = r | s_i = l] \). Let \( p \) denote this probability. Then we have \( \pi = \frac{q^2 + (1 - q)^2}{q^2 + (1 - q)^2 + 2q(1 - q)p} \), which means \( p = \frac{(q^2 + (1 - q)^2)q(1 - q)}{2q(1 - q)^2} \), and this expression simplifies to \( f(v) \).

\[\Box\]

## D Omitted proofs in Section 6

Suppose agent \( i \) has prior \( p_L \) and agent \( j \) has prior \( p_R \). Moreover suppose that all of agent \( i \)’s other neighbors honestly report their signal to \( i \) (e.g. they have the same prior as \( i \)). Would \( j \) honestly report her signal to \( i \) as well? The next lemma shows that there is an instance in which agent \( j \) benefits from persuading agent \( i \).

**Lemma D.1.** Suppose \( p_j[\theta = 1] < p_i[\theta = 1] \). Suppose \( j \) is not a like-minded neighbor, and \( i \) is not obstinate. Let \( S_i = \{s_i, \{s_k\}_{k \in N(i) \setminus j}\} \). Then there exists a realization of \( S_i \) such that \( p_j[\theta = 1 | s_j = r, S_i] \in (0, \frac{1}{2}) \) and \( p_i[\theta = 1 | s_j = r, S_i] \in (\frac{1}{2}, \frac{q^2 + (1 - q)^2}{q^2 + (1 - q)^2 + 2q(1 - q)p}) \).

**Proof.** Lemma D.1 directly follows from the definition of obstinate. Indeed if no such \( S_i \) exists, then either \( p_i[\theta = 1 | s_j = r, S_i] \) is always above \( \frac{q^2}{q^2 + (1 - q)^2} \), in which case \( i \) is obstinate, or whenever \( p_i[\theta = 1 | s_j = r, S_i] \in (\frac{1}{2}, \frac{q^2}{q^2 + (1 - q)^2}) \), we have \( p_j[\theta = 1 | s_j = r, S_i] \) also lie in the same interval, in which case \( i \) is a like-minded neighbor of \( i \). Note that
the fact that all other neighbors \( k \in N(i) \setminus j \) tell the truth is an important assumption. Without this assumption we must define “obstinate” with respect to other neighbors’ messages \( \{\tilde{s}_{ki}\}_{k \in N(i) \setminus j} \).

We now prove Lemma 6.3.

**Proof of Lemma 6.3.** When agents \( i \) and \( j \) have posteriors on opposite sides of \( \frac{1}{2} \) as in Figure 6, agent \( j \) has the incentive to persuade agent \( i \). More precisely, suppose all of \( i \)'s other neighbors \( k \in N(i) \setminus j \) tell her the truth. Then agent \( j \) chooses her reporting strategy as follows. First she finds the critical point at which she wants to persuade \( i \).

Let

\[
p^*_{ji} = \max_{S_i} p_i[\theta = 1|S_i]
\]

such that \( p_j[\theta = 1|s_j = r, S_i] \in (0, \frac{1}{2}) \) and \( p_i[\theta = 1|s_j = r, S_i] \in \left(\frac{1}{2}, \frac{q^2}{q^2 + (1-q)^2}\right) \). Whenever agent \( i \)'s interim belief \( p_i[\theta = 1|S_i] \) falls between \( \frac{1}{2} \) and \( p^*_{ji} \), agent \( j \) wants to persuade agent \( i \) to take action 0. For \( j \)'s persuasion strategy, we could directly apply Lemma 4.1. We have \( \sigma_{ji}(l) = l \) with probability 1, and \( \sigma_{ji}(r) = l \) with probability \( \min\{1, \frac{q-p^*_{ji}}{p^*_{ji}-(1-q)}\} \). Therefore agent \( j \) benefits from persuasion if other agents truthfully report their signals, and truthful reporting is not an equilibrium strategy for \( j \). \( \square \)

We next prove Proposition 6.5.

**Proof of Proposition 6.5.** Suppose agents \( i \) and \( j \) form a blocking pair, where \( p_i < \frac{1}{2} < p_j \). Let \( i' \) denote another agent with prior \( p_L \) and \( j' \) another agent with prior \( p_R \). Agents \( i \) and \( j \) both prefer the links \( (i, j), (i', j') \) over the links \( (i, i'), (j, j') \). Normalize everyone’s payoff in an assortative network to 0. In the deviated matching, agent \( i \)'s payoff is proportional to

\[
\min\{1, \frac{p^*_{ji} - (1-q)}{q - p^*_{ji}}\} - \min\{1, \frac{q - p^*_{ij}}{p^*_{ij} - (1-q)}\},
\]

and agent \( j \)'s payoff is proportional to

\[
\min\{1, \frac{q - p^*_{ij}}{p^*_{ij} - (1-q)}\} - \min\{1, \frac{p^*_{ji} - (1-q)}{q - p^*_{ji}}\}.
\]

We cannot have both agents \( i \) and \( j \) getting positive utility. Hence agents \( i \) and \( j \) cannot mutually prefer this deviation. As a result assortative network is stable. \( \square \)
E Omitted proofs in Section 7

Proof of Lemma 7.1. Since agents 1 and 3 have symmetric priors, their persuasion strategies are also symmetric. We have \( a_1(s_1, s_3) = a_3(\tilde{r}_1, r_3) \) whenever \( s_1 + r_3 = 1 \) and \( s_3 + r_1 = 1 \). Consequently we can rewrite (7.2) as follows:

\[
\int_{s_1, s_3 < \delta} p_1[1 - s_1, 1 - s_3] \cdot \mathbb{E}_{\theta|1-s_1,1-s_3,p_1}[-(\theta - a_3(s_1, s_3))^2 + (\theta - a_3(\tilde{s}_1, s_3))^2].
\]

This term is equal to

\[
\int_{s_1, s_3 < \delta} p_1[1 - s_1, 1 - s_3] \cdot \mathbb{P}[s_3(\tilde{s}_1, s_3) \neq a_3(s_1, s_3)] \cdot 2 \cdot \left| \frac{1}{2} - p_1[\theta = 1|1 - s_1, 1 - s_3] \right|.
\]

We have \( p_1[\theta = 1|1 - s_1, 1 - s_3] = p_1[\theta = 0|s_1, s_3] \). Consequently the loss is equal to

\[
\int_{s_1, s_3 < \delta} p_1[1 - s_1, 1 - s_3] \cdot \mathbb{P}[s_3(\tilde{s}_1, s_3) \neq a_3(s_1, s_3)] \cdot 2 \cdot \left| \frac{1}{2} - p_1[\theta = 0|s_1, s_3] \right|.
\]

In comparison the gain (7.1) is equal to

\[
\int_{s_1, s_3 < \delta} p_1[s_1, s_3] \cdot \mathbb{P}[s_3(\tilde{s}_1, s_3) \neq a_3(s_1, s_3)] \cdot 2 \cdot \left| \frac{1}{2} - p_1[\theta = 1|s_1, s_3] \right|.
\]

Note that the last term \( \left| \frac{1}{2} - p_1[\theta = 1|s_1, s_3] \right| \) is equal to \( \left| \frac{1}{2} - p_1[\theta = 0|s_1, s_3] \right| \). As a result the difference between the gain and the loss boils down to the first terms \( p_1[1 - s_1, 1 - s_3] \) and \( p_1[s_1, s_3] \). As \( \epsilon \to 0 \) we have \( p_1[1 - s_1, 1 - s_3] \to 0 \) and \( p_1[s_1, s_3] \to 1 \), so the gain exceeds the loss for small enough \( \epsilon \).