community. A community c is a group of people who have more dense connections within the group than to the rest of the people. For example, in Figure 4, the network forms two communities (shaded nodes 1, 2, 3, 4, 5 and unshaded nodes 6, 7, 8, 9) because the inter-connections within each community are more dense than to the intra-connections between two communities.

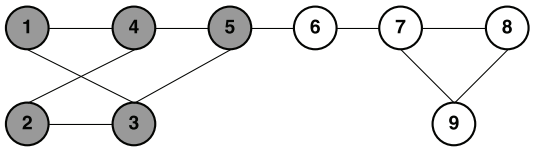

Figure 4: A social network with two communities.

The objective is to discover social communities $C = \{c_1, c_2, c_3, \ldots, c_k, \ldots, c_m\}$ based on both social network structure and opinion variations posted by each node u in the social graph, such that every node u in each detected community c_k not only closely connects to others, but also shares similar opinions on similar topics.

As social communities are generally more than one and are sometimes overlapping, we treat each community as a multinomial distribution over users. Thus, for each user, the conditional probability $P(u \mid c)$ measures how likely a user u belongs to the community c. The goal is, therefore, to find out the conditional probability of a user given each community.

Also, for each c_k in C, we give a concise description (community profile) with regard network structure, social properties, interested topics, and opinion summaries. In other words, the problem is how to produce an opinion summary, such that the summary could show what kind of group of social media users (common interests, interested topics) have what attitudes toward different topics about hotly talked objects on the social network.

3.2 POT: People Opinion Topic Model

To discover social communities based on users' opinion, we propose a graphic probabilistic generative model, People-Opinion-Topic (POT), which jointly models users' community, topic, and associated sentiment in a unified manner. It considers the formation of communities as a result of semantic similarity, opinion consistency, and network proximity among social media users.

Figure 5 shows the model structure of POT, and the relevant notations are listed in Table 1. Generally, users have multiple affiliations in the real world. Correspondingly, users in a social network also have multiple community memberships. We associate each user u with a community probability vector θ_c. A community c is assigned to a user u when u expresses opinion o on topic z.

Users within the same communities tend to have the same interests (topics) and share the same opinion orientation. Therefore, we associate each user a latent variable z generated from their interest distribution ϕ_u to indicate one's interested topic. Similarly, when expressing an interested topic z, a user is expressing his/her opinion o towards the topic z. Thus, we also associate each user's an opinion variable o from opinion distribution π_u. Note that in the traditional topic models such as LDA, a document contains a mixture of topics, and each word has a hidden topic label.

NOTATIONS	DESCRIPTION
D	The number of tweets
N	The number of tokens in a tweet T_d
M	The number of social communities
V	The number of users
u_v	v^{th} user
o_d	opinion orientation of a tweet T_d
z_d	topic of a tweet T_d
w_n^d	n^{th} word in tweet T_d
c_k^v	k^{th} community assigned to user u_v

Table 1: Notation of parameters

This is reasonable for long documents. However, the document D in twitter (a short text within 140 words limitation) is usually very short and is most likely to be about a single topic. Thus, in TOP model, all the words in D are assigned with a single topic z, and they are generated from the same word distribution ψ_z.

To easily integrate out $\lambda, \epsilon, \alpha, \psi, \beta$, we adopt conjugate priors in our model. Specifically, we place a Dirichlet prior over each multinomial distribution $(\theta, \Gamma, \psi, \phi)$, and a Beta distribution over the Bernoulli distribution π. The following distributions are drawn:

$$
\begin{array}{lcl}
\Omega \mid \sigma & \sim & Dirichlet(\sigma) \\
c_k \mid \Omega_v & \sim & Multinomial(\Omega_v) \\
\theta_c \mid e & \sim & Dirichlet(e) \\
u_k \mid \theta_c & \sim & Multinomial(\theta_c) \\
\pi_u \mid \beta & \sim & Beta(\beta) \\
o_d \mid \pi_u & \sim & Bernoulli(\pi_u) \\
\phi_u \mid \alpha & \sim & Dirichlet(\alpha) \\
z_d \mid \phi_u & \sim & Multinomial(\pi_u) \\
\psi_t \mid \lambda & \sim & Dirichlet(\lambda) \\
w_n^d \mid \psi_u & \sim & Multinomial(\psi_u)
\end{array}
$$

Based on the model, we obtain the joint distribution of the observed and hidden variables as described in Equation 1.

Consider a user u is a member of an opinion based community c and share common interests and opinions with others in the community. When he/she posts tweets on a specific topic z, he/she first selects the community membership c by his/her community's distribution θ_c. After choosing the community, he/she selects a topic z and opinion o, which are consistent with other group members. With the chosen topic z and opinion o, words a set of words is generated from the topic's word distribution. The generative process in POT model is summarized as following steps:

- Since a user may belong to different communities under the multinomial distribution, a user posts tweets following the atmosphere (similar opinion towards common topics) in their communities.
- When posting a tweet, the user firstly pick up a topic from a multinomial distribution and an opinion orientation from the binomial distribution.
- After the topic and opinion are decided, words are chosen from the topic according to a multinomial distribution.