The purpose of information source detection problem (or called rumor source detection) is to identify the source of information diffusion in networks based on available observations like the states of the nodes and the timestamps at which nodes adopted the information (or called infected). The solution of the problem can be used to answer a wide range of important questions in epidemiology, computer network security, etc. This dissertation studies the fundamental theory and the design of efficient and robust algorithms for the information source detection problem.

For tree networks, the maximum a posterior (MAP) estimator of the information source is derived under the independent cascades (IC) model with a complete snapshot and a Short-Fat Tree (SFT) algorithm is proposed for general networks based on the MAP estimator. Furthermore, the following possibility and impossibility results are established on the Erdos-Renyi (ER) random graph: (i) when the infection duration \( t_u < \frac{2}{3} t_u \), SFT identifies the source with probability one asymptotically, where \( t_u = \left\lceil \frac{\log n}{\log \mu} \right\rceil + 2 \) and \( \mu \) is the average node degree, (ii) when the infection duration > \( t_u \), the probability of identifying the source approaches zero asymptotically under any algorithm; and (iii) when infection duration < \( t_u \), the breadth-first search (BFS) tree starting from the source is a fat tree. Numerical experiments on tree networks, the ER random graphs and real world networks show that the SFT algorithm outperforms existing algorithms.

In practice, other than the nodes’ states, side information like partial timestamps may also be available. Such information provides important insights of the diffusion process. To utilize the partial timestamps, the information source detection problem is formulated as a ranking problem on graphs and two ranking algorithms, cost-based ranking (CR) and tree-based ranking (TR), are proposed. Extensive experimental evaluations of synthetic data of different diffusion models and real world data demonstrate the effectiveness and robustness of CR and TR compared with existing algorithms.