Chapter 1

Introduction

Peer-to-peer (p2p) paradigm for building distributed systems is becoming extremely popular as more and more novel applications (like VoIP, Instant Messaging, file sharing etc) are invented and successfully deployed [69]. Peer-to-peer system provides an architectural paradigm where every node performs both the role of server and client [27, 58]. They exchange information and services directly with each other without any hierarchical organization or centralized control. The main advantage of the p2p paradigm is that it allows the construction of systems of unprecedented size and robustness since all clients provide resources, including bandwidth, storage space, and computing power. Thus, as nodes arrive and demand on the system increases, the total capacity of the system also increases simultaneously. This is not true for a traditional client-server architecture, in which adding more clients could mean slower data transfer for all users. Because of these desirable qualities, many researchers have focused on understanding the issues surrounding the p2p networks and improving their performance.

Peers in the peer-to-peer networks are typically connected via ad hoc overlay connections. If a participating peer knows the location of another peer in the network, then a logical link may be established from the former node to the latter. The logical links among the peer nodes form the overlay network over the physical topology. The nature of connection of this overlay network determines whether p2p system is centralized, structured, or purely decentralized. Each such class of p2p systems has its own pros and cons.

The biggest advantage of pure decentralized p2p system (where peers randomly connect each other in a self organizing manner) is its robustness. However functions upon such system tend to be inefficient; for example, search in pure p2p networks amounts to flooding the network with query messages. The flooding mechanism generates large number of redundant query packets in the network which misutilizes the valuable bandwidth and makes the unstructured P2P systems being far from scalable. Superpeer network has proved to be the solution to such problem as it can combine the efficiency of client-server architecture with the autonomy, load balancing provided by the pure p2p networks. Hence, superpeer networks have emerged as the most dominant topology among the unstructured p2p networks. Most of the commercial systems like KaZaA, Skype use superpeer networks as the underlying architecture. In these systems, nodes are selected as superpeers on the basis of their larger capacity and greater capabilities from among the set of peers. Superpeer nodes containing higher bandwidth (hence connectivity) and resource connect to each other forming the upper level in the network hierarchy. Each superpeer works as a server on behalf of a set of client peers who form the lower level of network hierarchy [139,170]. Superpeer nodes route messages over the upper level of overlay network, and submit and answer queries on behalf of the pure peers and themselves. Hence, most of the query traffic flows through the superpeer layer (upper level) which in effect reduces the bandwidth consumption of the overall networks.

1.1 Formation and Dynamics of superpeer networks

The superpeer networks are formed and maintained as a result of several node and link dynamics like bootstrapping, peer churn, attack, link rewiring, upgradation of the peers to the superpeers etc. All these dynamics have a significant impact on the network topology as well as on the QoS of different p2p services like efficient search, file downloading etc. A brief description of these dynamics is given below. The detailed survey appears in Chapter 2.

Bootstrapping: The superpeer networks like Gnutella are formed mainly as a result of the bootstrapping protocols executed by peer servents like Limewire, Mutella, Gtk-Gnutella and Gnucleus. To join the network, incoming peers execute bootstrapping protocol through peer servents in which they discover other on-line peers in the network and send connection requests to them [34]. Bootstrapping protocols exploit physical properties of the online peers like resource content, processing power, storage space, connectivity etc. The protocols also take the finiteness of bandwidth of each online peer into consideration. At the time of joining, the incoming peer gets the list of online peers from web cache servers which are the distributed repositories for maintaining the information of 'good' online peers in the network [82]. These initial neighbor peers determine the new peer's location in the overall topology, and consequently its search and download performance. Hence, peers try (prefer) to join to 'good' (resourceful) nodes; all existing bootstrapping protocols are essentially directed towards fulfilling this basic objective [62, 104].

Peer churn and attacks: In superpeer networks, a peer joins the system when a user starts the servent, uses available resources of other peers (e.g., CPU, storage, bandwidth) while offering its own resources, and leaves the system when the user exits the application at some arbitrary later point in time. The independent arrival and departure by thousands or millions of peers create the collective effect of peer churn. Churn significantly affects both the design and evaluation of P2P systems, overlay structure [159] and the resiliency of the overlay [158]. In addition to that, important peers are also targeted for attack [138, 145]. Denial Of Service (DoS) attack [138] drown important peers in fastidious computation so that they fail to provide any service requested by other peers. Attackers mount more powerful attacks by leveraging the resources of multiple peers; these attacks are known as distributed denial of service (DDoS) attack [136]. Eclipse attack, Sybil attack, worm propagation, file poisoning, file pollution [26,145] are some of the important attacks that also affect the connectivity of the p2p networks. In summary, these peer churn and attacks cause serious threat to network resilience[1] as they have the potential of breaking down the connectivity among the peers in the network.

Link rewiring: Link rewiring is another internal dynamics that frequently occurs within the p2p networks. The peer node often disconnects the existing connection

[1] In this thesis, we do not differentiate between the terms stability and resilience. They are therefore used interchangeably.

with its neighboring nodes and establishes new connections with other 'good' online peers in the network. This rewiring operations, executed by peer servent, improve QoS of different p2p services by keeping 'resourceful' nodes as neighbors and play a major role in maintaining connectivity among the peer nodes specially in the challenged environment of churn and attack.

In summary, peer dynamics (churn and attack) disrupt the connectivity among the nodes, hence affect the stability of superpeer networks. In addition, bootstrapping and link rewiring play a major role in the formation and maintenance of the superpeer topology. The performance of network is affected by several topological parameters like amount of superpeer nodes, network connectivity, diameter etc. Hence, proper understanding of all these dynamics and their influence on various topological parameters will help in shaping the QoS of different p2p services.

1.2 Challenges in p2p networks and limitations of the classical approach

Significant amount of work has been done by the p2p research community in the development of efficient bootstrapping protocols [31, 62, 78, 130, 156, 169]. Several of these protocols assume the presence of a centralized bootstrap server [81] later modified to function just in presence of distributed GWebCache [34, 82, 146]. In addition, several random address probe based [31,61] and locality aware bootstrapping protocols [34] are developed to minimize the bootstrapping time and to reduce the redundant traffic in the underlay topology. Most of the works done in this field are directed towards designing of bootstrapping protocols to improve the performance of the p2p services. Such ad hoc improvements seem to have limited utility compared to the overhead they incur. Side by side, it is not obvious why bootstrapping of nodes and other local dynamics lead to the emergence of bimodal network, which appear in superpeer networks like Gnutella. It is interesting to note that these activities (joining, churn, rewiring) are completely driven by individual peer servents who perform them to optimize their own quality of service oblivious of the performance of the entire network. Hence, there is no explicit effort by either Gnutella client or server program towards formation of superpeer topology. The performance of the superpeer networks heavily depends upon the topological properties of the emerging networks [20,139,144, 170]. This includes the network diameter, amount of superpeers in the network, peer-superpeer ratio etc. Hence, regulating these topological properties and subsequently improving the performance of various p2p services will prove to be an *useful* step for p2p research community. Due to its decentralized nature of formation, controlling the topological structure of the superpeer network is not a trivial task. However, to the best of our knowledge, little work has been done to calculate these network parameters that emerge through the bootstrapping, churn and rewiring processes. Hence a considerable amount of research need to be directed towards modeling node and link dynamics and analytically understanding their exact impact on the topology of the emerging superpeer networks. Proper understanding of the various node and link dynamics and their impact on the emergence of superpeer networks may provide some important insights to the network engineers for improving the quality of various p2p services.

Peer-to-peer networks also suffer from high rate of peer churn due to the continuous joining and leaving of the nodes in the network. In addition, stability of the network can get affected through intentional attacks targeted towards important peers [138, 145]. These dynamics of the peers often partition the network into smaller fragments which result in breakdown of communication among peers. We find that although several attacks and defence techniques are discussed in the literature, less attention have been paid to analyze the impact of such attacks upon the overall topology of the p2p networks. It is very important to maintain the connectivity of p2p network in order to sustain the regular p2p activities. Measurement based study and experimental analysis of resilience of p2p networks have been done by various researchers [146, 158, 159]. Some simulation based studies have also proposed design guidelines to construct robust p2p networks [130]. However, apart from simulation and experiment based study, stability analysis of the peer-to-peer networks also need to be undertaken from a theoretical perspective. More specifically for superpeer networks, design engineers often face the essential questions like, what is a good ratio of peers to super-peers in the network? How should superpeers connect to each other and with regular peers? How different topological parameters may affect the network stability against various node dynamics? How does size of the network play a role in determining stability? In summary, we can say that a comprehensive theory for understanding the stability of finite sized networks under any type of node disturbance is desirable. In this thesis, we try to address some aspects of these issues related to the stability and emergence of superpeer networks.

1.2.1 Complex network as a toolbox

The commercial peer-to-peer networks are quite large in size and contain millions of nodes. As discussed in the previous section, these large scale p2p networks are formed and maintained as a result of various node and link dynamics. P2p networks at any instant of time can be viewed as very large scale dynamic graph. However, it is difficult to apply traditional graph theoretic approaches for analyzing the properties of such large scale networks which are in a constant state of flux. Hence, the behavior of these systems can only be analyzed by observing various statistical properties of the network especially by applying the theories of network science. Large scale dynamic networks are found in various fields of study like sociology (social network, friendship network, film actors network), biology (protein-protein interaction network, metabolic network), linguistics (word co-occurrence network), information science (citation network), electrical technology (power grid, electronic circuits), computer science (internet, world wide web network) etc. Network theoretic approaches are widely used in analyzing these social, biological, and technological networks which display non-trivial topological features like heavy tail in the degree distribution, a high clustering coefficient, assortativity or disassortativity among vertices, community structure, and hierarchical structure. Large scale p2p networks can also be modeled as complex graphs and various theories related to network science may be applied in analyzing the behavior of p2p networks. Significant amount of work has been done in the field of complex networks in understanding the growth of complex networks in face of various node dynamics [11, 16, 18, 45, 87]. The basic assumption of all these works has been that a node joins the network based on preferential attachment, that is a new node generally attach itself to 'important' existing nodes. It has been widely seen that such behavior leads to the emergence of scale free or power law networks. The theoretical analysis of the formation and dissolution of giant component against random failures and attacks in large scale networks are mostly based upon the percolation theory and are discussed in [9, 28, 29, 127].

In this thesis, we utilize various concepts of complex network theory like percolation theory, continuum theory, etc and suitably modify them to analyze the dynamics of superpeer networks. The main contribution of the thesis is two folds;

1. Analyzing the stability of arbitrary size superpeer networks in face of peer churn and attacks,
2. Formal understanding of the emergence of superpeer networks in face of various local events like churn, link rewiring etc.

1.2.2 Objectives of the thesis

The principal objective of the thesis is to develop analytical frameworks for understanding the various dynamics in large scale dynamic superpeer networks. We primarily focus on two major topological properties, namely resilience and emergence of superpeer networks. Specific problems are :

- Development of an analytical framework to measure network stability against node dynamics like peer churn and attack.

  One of the main objectives of this thesis is to build up a complete analytical framework whereby given a topology and an attack scenario, one should be able to predict the exact point of breakdown of the network. Such a framework should also be able to explain the observed topological characteristics of superpeer networks. The effect of peer-superpeer degrees (and their respective fractions) on the network stability need to be illustrated. The network gets deformed after removal of a fraction of nodes along with their adjacent links due to node dynamics. The developed framework need to also precisely describe the topology of the deformed network after churn or attacks.

- Modeling bootstrapping protocols as node attachment rules and formally explaining the emergence of superpeer networks.

  Node attachment rules may be influenced by factors like shared resources, processing power, bandwidth etc. These abstract parameters need to be quantitatively represented and their impact on the topological properties of the superpeer networks need to be analyzed through such a growth framework. The framework should also be able to illustrate the impact of peer churn and rewiring on the network properties like amount of superpeer nodes, network connectivity, component sizes, network diameter etc. All these parameters have significant influence upon the quality of different p2p services. Hence the final objective of this thesis is to build up a comprehensive framework encompassing growth, node churn and link rewiring.

Keeping these above broad objectives in mind, the particular work done is outlined in the next section.

1.3 Contribution of the thesis

In this thesis, we develop theoretical frameworks to analyze the resilience and emergence of superpeer networks against several node and link dynamics. These frameworks are validated through simulation as well as real world data of Gnutella. We also discover several nonintuitive, interesting properties related to network topology that may be useful to the network researchers to improve the performance of the p2p services. The specific contributions are given below.

- **Stability analysis against peer churn and attacks**
  We model the superpeer networks with the help of degree distribution p_k (probability of a node of degree k) and peer dynamics by another probability distribution f_k (probability of removal of a node of degree k). We derive a critical condition for the stability of superpeer networks which undergo node dynamics. The degree distribution of the deformed network after node removal is

also calculated. The results obtained from the theoretical analysis are validated through stochastic simulation as well as by real data of Gnutella network. We measure the impact of fraction of superpeers in the network as well as their connectivity upon the stability of the network. The influence of network size as well as degree-degree correlation present in the real world networks like Gnutella is also analyzed.

- **Generalized model of node dynamics**
  We characterize the node dynamics as the various kinds of node removal processes. We view **degree dependent attack** as a broad class of node removal process which is able to capture peer churn and attacks. In this node removal strategy, the probability of removal of a node (f_k) having degree k is proportional to k^γ. We show that, by varying the attack parameter γ, we can generate the wide range of node dynamics, from random failure to deterministic attack.

- **Emergence of superpeer networks due to bootstrapping**
  We model bootstrapping protocols through node attachment rules. We show that a significant class of bootstrapping protocols may be viewed as a node attachment rule where the probability of joining of an incoming peer to an online node is proportional to the node property (shared resource, processing power, bandwidth) and degree of the online node. We identify that in p2p networks, bandwidth of a node is finite which restricts its maximum degree. A node, after reaching its maximum degree, rejects any further connection requests from incoming peers. We develop a formalism that calculates the degree distribution of an emerging superpeer networks based upon such bootstrapping process and bandwidth constraint. The proposed growth framework reveals that the interplay of finite bandwidth with node property plays a key role in the accumulation of superpeer nodes in the network. As an application study, we show that our framework, with some modification, can explain the topological configuration of commercial Gnutella networks.

- **Emergence of superpeer networks against peer churn and link rewiring**
  We refine the above growth framework where emergence of the superpeer networks is driven by the (a) joining of incoming nodes (b) random departure of peers due to peer churn and (c) rewiring of the existing links thereby biasing

connections towards resourceful peers. The analytical framework calculates a critical churn rate, upto which the qualitative nature of superpeers is preserved. It also discovers that in presence of proper rewiring, the QoS of p2p network shows graceful degradation in face of churn. Our theoretical model provides some empirical estimation of the node properties, churn and rewiring rate of the Gnutella network which is consistent with the measurement results.

1.4 Organization of the thesis

The organization of rest of the thesis is as follows. Prior to dealing with the proposed work, we report a survey on related research topics in Chapter 2. Chapter 3 focuses on modeling of superpeer networks as well as various node dynamics and analyzes the stability of superpeer networks against peer churn. In Chapter 4, we analyze network stability and topological deformation against attacks. Chapter 5 develops a formal framework to analyze the emergence of superpeer network from bootstrapping by incoming peers. In Chapter 6, we extend the framework to include peer churn and link rewiring. Chapter 7 concludes the thesis.