Phi Nguyen
Annotated Bibliography

Topic 1: Peer-to-peer chat

Short description: This app allows two(or more) people to instantly and directly connect with each other and exchange messages without any third party authorization. It operates completely on a peer-to-peer network, established among the users.

1. Hu, Zhou. "NAT traversal techniques and peer-to-peer applications." HUT T-110.551 Seminar on Internetworking. 2005. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.103.1659&rep=rep1&type=pdf
   The paper addresses the challenge that NAT(Network Address Translation) poses when dealing with Peer-to-peer communication. Although NAT is useful in a sense that it blocks unwanted incoming requests from the outside world and only allows some specific authorized requests, it is an obstacle when one wants to send some data to its peer. It gives four methods of NAT Traversal techniques:
   + Universal Plug and Play, developed by Microsoft, and only used for Windows computers, however it has been proved with insecurities problems.
   + Simple Traversal UDP Through Network Address Translators, a protocol that helps peers to send STUN requests and responses. There is a STUN server that sits between the requests to help the peers identify if it is behind a NAT. However, this method does not work with symmetric NAT
   + Application Level Gateway, embeds the IP addresses and information right in the packet, could be seen as a NAT extension component. Limitation is it often requires modification of NAT device(router), which is hard to deploy
   + UDP/TCP Hole Punching, a general and robust technique to establish the connection between two hosts. It uses a relay server that introduces the two hosts together, both or either could sit behind a NAT. The disadvantage is that it turns peer-to-peer into a Client-server communication(at the start of the communication), not purely peer-to-peer.

2. Stutzbach, Daniel, and Reza Rejaie. "Understanding churn in peer-to-peer networks." Proceedings of the 6th ACM SIGCOMM conference on Internet measurement. 2006. http://ix.cs.uoregon.edu/~reza/PUB/imc06-churn.pdf
   The paper introduces the definition and addresses the possible improvement of churn in P2P Network. It bases its study on different existing P2P systems such as BitTorrent. It also poses challenges in dealing with P2P in an untrusted world. Some notable difficulties that is directly related to my topic of research is
   + Missing data: in an unreliable data transmission session between two peers, how to handle missing data

+ Handling long session: in an unreliable transfer protocol between two peers, how to handle long session, in other words, how to maintain reliable connection when one of the two has weaker signal for example.
+ NAT: as illustrated in the first annotation, machines that sit behind a NAT cannot be reached from the outside world by normal protocol
+ Dynamic Addresses: DHCP and PPP dynamically assigns IP addresses to hosts. It remains a single IP address throughout the session, but once it restarts the session, it could appear with a different IP Address

3. Wang, Yao, and Julita Vassileva. "Trust and reputation model in peer-to-peer networks." Proceedings Third International Conference on Peer-to-Peer Computing (P2P2003). IEEE, 2003.
   https://d1wqtxts1xzle7.cloudfront.net/33965069/120_wang_y.pdf
   This paper addresses trust (peer's belief in another peer), and reputation (could be a third party centralized computing the trustworthiness of the peer). It is related to my topic of research in a sense that we could exclude potential unreliable peers such as non-existing peers, or peers that pose insecure threats, etc.
   Although this paper deals with file sharing peer-to-peer systems, one could adopt the techniques to a peer-to-peer chat system. The paper uses a Bayesian network-trust model, in which Download Speed, File Types, File Quality are the three main criteria in determining a trustworthiness of a file provider. The paper also introduces a recommendation system, where it asks other peers about the trustworthiness of a peer when it could not decide whether it should trust this peer (file provider) or not.

4. Halkes, Gertjan, and Johan Pouwelse. "UDP NAT and Firewall Puncturing in the Wild." International Conference on Research in Networking. Springer, Berlin, Heidelberg, 2011.
   https://link.springer.com/content/pdf/10.1007/978-3-642-20798-3_1.pdf
   This paper also deals with the challenge of NAT in a Peer to Peer Network, specifically with UDP transfer protocol.
   The authors approach the problem with a rather bottom up method. They first ask the question: what types of NAT are there, and how different/popular they are. It turns out that each NAT has a different success rate when communicating peer-to-peer, with a rather large standard deviation. However, this paper did not produce any solution to the problem, rather it only addresses the existing issues, and claimed that 75% of machines on the internet are not connectable. It also introduces some potential solutions in the future, such as the emergence of IPv6, however, they are quite tangible and without any proof.

5. Eppinger, Jeffrey L. TCP connections for P2P apps: A software approach to solving the NAT problem. Vol. 6. Technical Report CMUISRI-05-104, Carnegie Mellon University, 2005.
   http://reports-archive.adm.cs.cmu.edu/anon/anon/usr0/ftp/usr/ftp/isri2005/CMU-ISRI-05-104.pdf

This paper also addresses the problem with NAT in Peer-to-peer communication. It first introduces current approaches and limitations, some of which are listed above, with some additional examples of Skype and BitTorrent. It differs in other papers in which it clearly introduces the requirements for a system to be successful and designs their system accordingly. According to the paper, UDP is more favorable generally, because it does not need many requests to maintain the session between two peers. It also adopts the techniques of the relay server, but changes the name to "connection brokers". It can identify whether a peer is behind a NAT by letting the peer exchange network information and if the private address is the same as the public address then it is not behind a NAT. It uses an additional trick here: they use the same IP addresses and ports for communication with the broker and for the communication between the two peers.

6. Ford, Bryan, Pyda Srisuresh, and Dan Kegel. "Peer-to-Peer Communication Across Network Address Translators." USENIX Annual Technical Conference, General Track. 2005.
   https://www.usenix.org/legacy/event/usenix05/tech/general/full_papers/ford/ford_html/
   This is a very thorough paper on NAT Traversal with focus on NAT Hole Punching. Although the NAT devices were self-selected - they relied on people reporting their NAT configuration as their data - the result was very encouraging. In essence, to "punch a hole" in the NAT, we still need a relay server sitting between the connection of the two peers. However, there are two ways this connection can be initiated/maintained. The first one is that the server is just another "hop" between their connections, meaning it merely transfers what A wants to send to B and vice versa. This is proved to be very robust and reliable. However, we still need to rely on a server for the transmission of private data. Another way is the server becomes the waiter per se. When A wants to connect to B, A tells the server it wants to "book a table", and server keeps the private port of A. When B comes and tells its private port to the server and tells that it wants to connect with A, the server (or waiter) now introduces them to each other by forwarding the private IP address to each other. This can work with majority of NAT, and can still (somewhat) work even when one (or both) of the peers is behind multiple NATs.

Topic 2: Web-scraping/ Information Extraction

A web scraper that works on a high level (unlike language-dependent library where it extracts based on HTML elements). Users request general information such as people's information, job's description, statistics and it will search for those in targeted web pages.

1. - Cowie, Jim, and Wendy Lehnert. "Information extraction." Communications of the ACM 39.1 (1996): 80-91.
   http://staffwww.dcs.shef.ac.uk/people/Y.Wilks/papers/infoext.pdf
   This is a rather long paper that addresses the current state of the art in informational extraction, i.e extracting the most essential (targetted) information given a text. It has a direct relationship with my topic of research as it crawls the essential information from a

(html) web page. The paper is an introductory text to information extraction. It employs (mostly) Natural Language Processing(NLP) Techniques to evaluate a piece of text. The paper also introduces various current attempts to create a generic IE System, some can be helpful if this topic is chosen as the final topic
+ Filter: cross out irrelevant sentences
+ a Text Zoner: turns text into set of segments
+ Preprocessor: turns text into sequence of sentences, each of which being a sequence of lexical terms

2. - Grishman, Ralph. "Information extraction: Techniques and challenges." International summer school on information extraction. Springer, Berlin, Heidelberg, 1997.
https://cse.buffalo.edu/~regan/Talks/ie-survey-frascati-97.pdf
This talk is also an introductory text to information extraction. It differs from the previous paper that instead of focusing on the current employed techniques, it also emphasizes the importance of information extraction and its place in the current big data world. It also uses a different approach from the previous paper. Instead of categorizing the techniques generally, it introduces different currently interesting problems, such as name recognition, pattern matching, etc. Furthermore, it also introduces different challenges when coming to decide the design for the extraction machine such as when to parse (syntactically) and when not to parse, portability, etc. The paper also addresses (shortly) the importance of performance in designing the extraction algorithm.

3. Chang, Chia-Hui, et al. "A survey of web information extraction systems." IEEE transactions on knowledge and data engineering 18.10 (2006): 1411-1428.
https://d1wqtxts1xzle7.cloudfront.net/46008560/A_Survey_of_Web_Information_Extraction_S20160527-18159-1fxuimz.pdf
This paper targets a subset of the last two papers, one that will be the focus of my topic of research: web information extraction. It addresses three criteria when judging/comparing IE systems: task difficulties, techniques used, and automation degree. Different from two previous papers, this one drills the techniques that are used only for the web, however, with a rather large scope: from manually crafting the software to using unsupervised machine learning. It seems that any of the two ends would be very difficult to maintain and scale, a supervised (with closely manual support) is favored.

4. - Etzioni, Oren, et al. "Open information extraction from the web." Communications of the ACM 51.12 (2008): 68-74.
https://www.aaai.org/Papers/IJCAI/2007/IJCAI07-429.pdf
This paper is similar to the previous paper in a sense that it also focuses on the web information extraction subset and introduces various similar techniques that are currently used in the research community. The difference here is that instead of only listing current popular techniques, it guides the reader through a software called TEXTRUNNER. This software does not only extract information, but also compares it with other texts to determine the trustworthiness of a text. It also acts as a fact checker when judging a

piece of text. The result in this paper might be distant from my focus, however, there are some interesting techniques that can be employed in my research such as the method they use in building their learner, i.e self-supervised, single-pass extractor, redundancy-based accessor.