Thoughts on Human Emotions, Breakthroughs in Communication, and the Next Generation of Data Mining

Hillol Kargupta
University of Maryland, Baltimore County and Agnik

Abstract: *This paper revisits some of the breakthroughs in the field of communication that changed the human civilization and tries to understand where we are going tomorrow. It argues that while we have become very good in quickly connecting an entity with another entity world-wide as long as the former knows the address of the latter, current technology for taking the message or service from one individual to a large population of willing and interested individuals is still very primitive. We need technology for dealing with the new world of attention-economics. The paper also argues the currently popular centralized client-server model for the Internet applications may not work well in doing so. It offers some alternate thoughts borrowed from the nature and some emerging scalable applications that may offer some directions.*

Introduction

Communication played a critical role in the development of human civilization. We cry, laugh, smile, talk, and write. Each of these dimensions makes what we are and also help shaping what others are. We need others to shape our life and the vice versa. This fundamental need of our inner self has played a key role in motivating the development of the communication technology since the days of cave dwelling homosapiens. Now that we have reached the era of almost instant communication through the Internet, cell-phones, wireless networks, mobile ad hoc networks, (MANET) and vehicular networks (VANET), what else we need to further explore what we are and how we interact.

This paper argues that while we have become very good in almost instantly connecting an entity with another entity as long as the former knows the address of the latter, we have made little progress in taking the messages and services from one to a large population of willing and interested parties. The paper also argues that the current client-server model of communication in the Internet applications and social networking sites may not scale very well in connecting individuals over the next generation of the Internet using wired, wireless, and ad hoc networks

Section 1 discusses some of the early breakthrough in communication technology and their local nature. Section 2 revisits some of the early efforts in expanding the range of local communication. Sections 3 and 4 identify some of the challenges for this age of Internet-era communication. Section 5 makes a note of some lessons from the nature and discusses them on the ground of scalable local distributed algorithms. Section 6 identifies the role of data mining in the alternate communication architecture that attempts to bring people closer at a global range through local interaction in a scalable manner with less reliance on centralized controls.

1. The First Breakthrough: Speech and Local Communication

One of the biggest breakthroughs in the history of mankind was the evolution of language for communication. It is believed that the early forms of language evolved about 200,000 years ago in homosapiens. This gave us the capability to communicate among a relatively small group of individuals when they are in proximity. We could now smile, laugh, cry, and also talk for sharing different facets of our life with others. There is no doubt that speech revolutionized human civilization and played a key role in where we are today. Evolution of speech gave us a tool with a constraint----local proximity. We can talk to others; but in absence of any other technical aid we can communicate with only those who are physically nearby and can listen to what we are saying.

Figure 1. Cave painting for local communication.

However, this was not sufficient. As relatively complex social-structure developed, people felt the need for communicating with others who are at a remote place either spatially or temporally. Simple verbal communication was not sufficient. Then language started shaping up in the written form. Cave paintings started appearing around 30,000 BC (e.g. Chauvet Cave in Southern France). Petroglyphs, Pictograms, and Ideograms started emerging in different societies. This lead to the invention of the first writing systems in the late 4th millennium BC. This allowed documentation of events and communications in a more permanent form. People learnt how to communicate over time. For example, a cave dweller could then observe a hunting event and document that for the posterior generations. Although this allowed us to reach a bigger audience over time, the spatial locality constraint still remained. If a cave dweller had a message for you on the walls of a cave, you did have to go there in order to retrieve the message.

2. Distance Communication: One to Few More

Mankind tried to invent techniqes for removing the restriction on spatial locality for communication. For example, smoke signals, fire beacons, heliographs were used in early days for expanding the scope of the spatial locaility. Courier-based postal systems emerged in Egypt during 2400 BC. Iran, India, and China are some of the places where matured postal systems were developed during 500-185 BC. The postal system gradually offered access to distance communication to common people. It made communication a lot more convenient. You can write a letter with an address, drop it at a fixed location, and it will most likely reach the destination if you are willing to pay for the stamps. The main constraint is that the postal system takes time, it is not very personalized, and you need the address before you can communicate with someone.

We wanted more convenience. We wanted instant distance communication from the convenience of our living room in our house or the workplace. Commercial telegraph system was invented in 1837. The telephone systems appeared in 1871. Around the same time radio was invented for wireless communication. All these technologies greatly enhanced the convenience of distance communication. It also gave the receiver some control. If you do not want to participate then you can opt out (e.g. hang up the phone, throw away the mail that you received from the post office). However, some of the fundamental constraint remained there:

1) You still needed to know the address of the destination in order to communicate with someone at a distance location.
2) It was still very hard to reach a very large number of people all over the world who are interested in listening to you message.

3. The Internet Era

The 20th century offered us the computer-based communication technology. The Internet further enhanced the convenience factors of the distance communication. We can reach an even larger number of individuals all over the world; communication is cheaper; it is fast and multi-media friendly. However, the fundamental constraints are still prevalent. You need to know the address of the computer node where you are sending the message and it is still hard to get your message out there to a large interested audience. The current Internet-era solutions for these problems are fundamentally similar to what we have in the postal system.

One of those is sending millions of junk mails to a large collection of addresses. Fundamentally this is very similar to the junk-mails that we receive in our postal mailbox, only in a larger scale. This is a fairly primitive concept and unlikely to scale producing a large number of satisfactory clients. In fact most individuals view spams as an unwelcome mode of communication.

Another emerging mechanism is based on the so called social-networking web-sites. These web-sites allow you to sign-up, post personal information and other content. Others can search and browse your information. If there is a match of interest then the site typically offers a mechanism to connect them. This mechanism is gaining popularity among different sections of the society in absence of any better solutions for address-free communication with a large number of individuals all over the world and making your message heard by a lot of interested people. However, this approach has several problems. The following section will discuss those.

4. The Missing Piece of Puzzle: From One to Many

Current client-server models for social networking type infrastructures have several problems:

1) **Economics of Mass Communication:** Often these sites have business interests that are driven by the economics. That means if a match-making or a social communication creates value for the business then eventually those are the ones that will be promoted by the site. For example, these days many news-websites allow the readers to send images and make news. This is similar to the readers' columns in good old newspapers. The important thing to note is that they moderate this news. The owner/editor of the web-site decides which one of your images if any gets posted. If you have a message for the world that does not help the economics of the news web-site then it is unlikely to be used.

Figure 2. An ant colony in the Pirin Mountains, Bulgaria.

2) **Privacy and Intellectual Property Issues:** Many of these sites want you to publish your content to their website. There are contents from many aspects of life that you may want to control because of privacy or intellectual property issues. Existing business models basically expect you to trust the owner of the site with your content.

3) **Not Scalable:** The biggest problem with the current centralized approach is that it is not very scalable. This is particularly troublesome for the next generation of Internet based on wired, bandwidth-constrained wireless, and ad hoc networks. The approach is based on a centralized solution for a fundamentally large distributed environment. It is equivalent to saying that every node in the network must communicate with a single node and let that node handle all the data processing tasks. If I have a message that should reach a large audience of interested individuals and I do not have the addresses of these individuals then the site must be able to analyze content of the message and find the matches with the interest profiles of individuals. Doing this may

require performing various data mining tasks such as clustering, indexing, and classification among others. As the volume of data and the size of the Internet increases along with heterogeneity (e.g. wired and wireless) it will be harder to scale such a centralized approach. This will particularly be harder in the mobile wireless world.

5. Any Better Approach?

While we do not yet have an existing solution that can take someone's message to a large appropriate audience in a more efficient manner, we may have some clues toward alternate approaches. This section briefly outlines some of those.

Figure 3. Swarm behavior in fish schools.

Nature offers many scalable, large complex systems where a large number of entities interact with the others they need to in an efficient manner. Individuals become part of a peer-to-peer network where messages from one entity get transformed and transmitted to other entities in the network. For example, consider an ant colony. Large ant colonies are known to have hundreds of millions of ants. In a colony, ants locally interact with each other using a set of simple rules and the global behavior emerges out this local behavior despite the lack of centralized control. In fact human civilization is also an example of such behavior. Different parts of the world produced different societies and different cultures based on there local rules of interactions. The global pattern of behavior for the human race is indeed a result of such local interaction and some global norms. The swarm behavior of large complex distributed systems has been noticed in many natural systems such as schools of fishes, migratory birds, termite colonies, and many other species.

Global communication through local interactions also appears to have a strong theoretical basis. [1][2]. The literature on distributed algorithms and systems point out that synchronized communication through a single node does not usually produce reliable efficient algorithms for distributed algorithms. Local algorithms that work by bounding the communication cost of each node to a reasonable amount often scale much better compared to a centralized solution in a large asynchronous distributed environment.

6. Global Communication through Local Interactions & Distributed Data Mining

In order to take someone's message, finding the appropriate audience, and delivering that worldwide in a scalable manner would require developing technology for the following key problems: (1) Less reliance upon single-entity owned centralized client-server model of computation with more emphasis decentralized emergence on global behavior through local interactions; (2) privacy-sensitive content analysis and match-making between the source and the interested parties in a distributed decentralized environment.

Although we have a long way to go in solving these problems, we are starting to see some possible directions. The methodology for achieving global communication through efficient but strictly local interactions is drawing attention. For example, peer-to-peer (P2P) networks have been gaining popularity in many domains. P2P systems work by using the computing power, storage, and bandwidth of the participants of a network. Unlike client-server systems, P2P systems do not rely upon the servers to carry out most of the computation and storage-intensive tasks. P2P systems such as Gnutella, Napster, e-Mule, Kazaa, and Freenet are increasingly becoming popular for many applications that go beyond downloading music without paying for it. Examples include P2P systems for network storage, web caching, bio-informatics, astronomy, searching and indexing of relevant documents and distributed network-threat analysis.

Matchmaking and personalized information retrieval in such P2P environments would require distributed data clustering, indexing, and classification algorithms that work in a decentralized communication efficient manner. P2P distributed data mining algorithms [3][4] offer many interesting applications such as client-side P2P Web-mining. Many popular Web servers use web-mining applications to analyze and track users' click-stream behavior. Now imagine client-side web mining that does the same for Web site visitors (rather than host servers) by analyzing the browsing histories of many users connected via a P2P network. Today, site visitors have no direct access to the results of Web mining algorithms running on the servers, but a client-side P2P Web-mining system [5] could empower visitors with click- stream data mining for advanced applications such as P2P search, interest-community formation, and P2P electronic commerce.

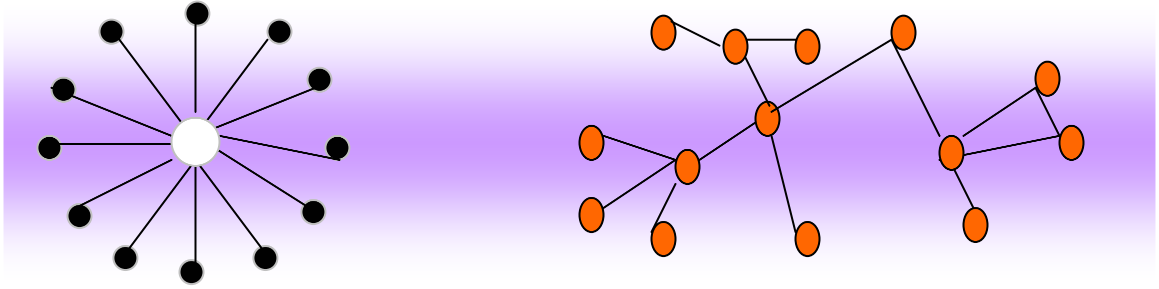

Figure 4. (Left) A centralized communication through a hub-and-spoke architecture. (Right) Global communication through local control and interaction.

For example an application like this may be able to find the best deal in the cell-phone market by analyzing the search history of multiple users in a privacy-preserving manner. Clearly, maintaining users' privacy will be an important issue and the field of privacy-preserving distributed data mining may offer some solutions. Similar methodology can be used for matchmaking and finding consumers of other services or messages in a decentralized, distributed P2P-like environment. P2P News, P2P e-commerce, and P2P exploratory astronomy are some examples. We need more work along these directions for truly bringing the power to the people---linking producers of messages and services to interested consumers without relying too much upon a centralized entity while protecting the privacy of the involved parties. This should be a key focus for the next generation of data mining research.

References

1. D. Peleg. (2000) Distributed Computing: A Locality-Sensitive Approach, SIAM,Philadelphia.

2. M. Naor and L. Stockmeyer. (1995). What can be computed locally? SIAM Journal on Computing, Volume 24 , Issue 6, Pages: 1259 - 1277

3. H. Kargupta and K. Sivakumar, (2004) Existential Pleasures of Distributed Data Mining. Data Mining: Next Generation Challenges and Future Directions. Editors: H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha. AAAI/MIT Press.

4. S. Datta, K. Bhaduri, C. Giannella, R. Wolff, and H. Kargupta. (2006). Distributed Data Mining in Peer-to-Peer Networks. (Invited submission to the IEEE Internet Computing special issue on Distributed Data Mining), Volume 10, Number 4, Pages 18 - 26.

5. K. Liu, K Bhaduri, K. Das, P. Nguyen, H. Kargupta (2006). Client-side Web Mining for Community Formation in Peer-to-Peer Environments. ACM SIGKDD Explorations. Volume 8, Issue 2, Pages 11 - 20.