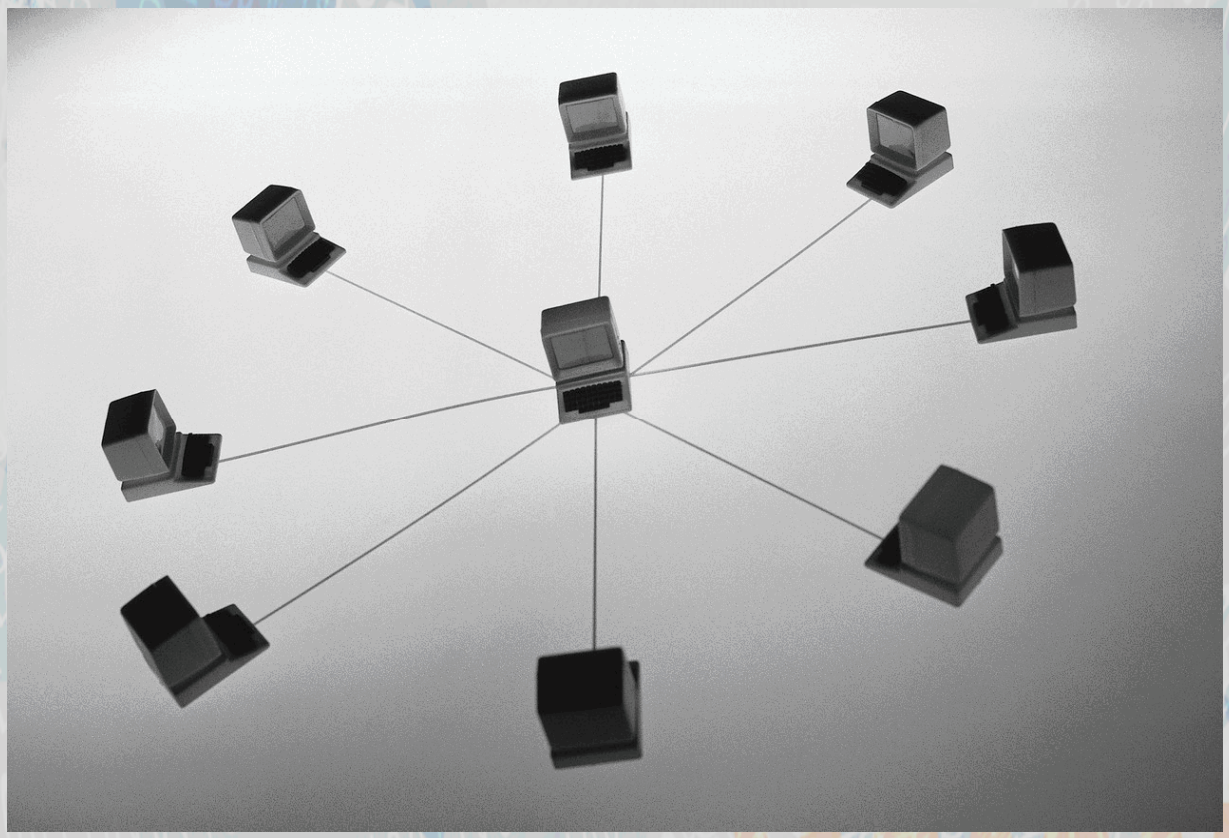
CSRS 2007

Proceedings of the 1st Villanova University Undergraduate Computer Science Research Symposium



December 10 & 12, 2007
Computing Research Topics

Welcome from the Conference Chairs

Welcome to the 1st Villanova University Undergraduate Computer Science Research Symposium. This Fall 2007 symposium is the culmination of the research efforts of nine dedicated undergraduate students in the Computing Research Topics course at Villanova University. These students spent the past four months exploring the academic research process, including selecting a research topic and performing a literature survey, then writing an abstract, extended abstract, literature review and research proposal, and finally preparing a conference presentation.

The topics in this symposium cover a wide range of computer science ideas, from networks and security to facial recognition, speech recognition and parallel computing. The papers contained in these proceedings each present a concise survey of the subjects they cover, followed by a proposed direction for future research. Each of these works will be considered for submission to the Villanova University Undergraduate Research Grant Program, where student researchers are awarded special funding to advance their proposed research topic.

We are truly proud of the accomplishments of these nine outstanding computer scientists, and hope you enjoy reading the following collection of their persistence.

Most sincerely,

Mirela Damian, Ph.D.
Thomas Way, Ph.D.
Conference Chairs
Computing Research Topics professors

Perhaps the most valuable result of all education is the ability to make yourself do the thing you have to do, when it ought to be done, whether you like it or not.
- Thomas Henry Huxley

A journey of a thousand miles must begin with a single step
- Lao-Tzu

Proceedings of the 1st Villanova University Undergraduate Computer Science Research Symposium

Encryption: Strengths and Weaknesses of Public-key Cryptography

Matt Blumenthal
Department of Computing Sciences
Villanova University, Villanova, PA 19085
CSC 3990 – Computing Research Topics
matthew.blumenthal@villanova.edu

Abstract

Public key cryptography has become an important means of ensuring confidentiality, notably through its use of key distribution, where users seeking private communication exchange encryption keys. It also features digital signatures which allow users to sign keys to verify their identities. This research presents the innovations in the field of public-key cryptography while also analyzing their shortcomings. We present methods of improving upon the weaknesses that include techniques involving double encryption and mutual authentication. These contributions introduce new levels of security to the subject with ideas to combat man in the middle attacks and other hacker scenarios. Public-key encryption with digital signatures offers both security and data integrity against most attackers.

1. Introduction

Public key cryptography has become an important means of ensuring confidentiality, notably through its use of key distribution. Key distribution is an approach where users seeking private communication exchange encryption keys, while digital signatures allow users to sign keys to verify their identities. This research explores the strengths and weaknesses of public key cryptography, examining potential flaws and methods of correcting them.

2. Secret-key Cryptography

Secret-key cryptography, also known as symmetric-key cryptography, employs identical private keys for users, while they also hold unique public keys. "Symmetric-key" refers to the identical private keys shared by users. Users employ public keys for the encryption of data, while the private keys serve a necessary purpose in the decryption of data. People wishing to engage in a secure exchange of information will swap public keys and use some method to ensure the existence of identical private keys. In theory,

private keys would be brought into the transaction through either the duplication of an existing key or the creation of two identical keys. In modern practice, users utilize key generators to create both keys, but the private keys must still be distributed in a confidential mode.

2.1 Strengths

The private keys used in symmetric-key cryptography are robustly resistant to brute force attacks. While only the one-time pad, which combines plaintext with a random key, holds secure in the face of any attacker regardless of time and computing power, symmetric-key algorithms are generally more difficult to crack than their public-key counterparts. Additionally, secret-key algorithms require less computing power to be created than equivalent private keys in public-key cryptography. [3]

2.2 Weaknesses

The biggest obstacle in successfully deploying a symmetric-key algorithm is the necessity for a proper exchange of private keys. This transaction must be completed in a secure manner. In the past, this would often have to be done through some type of face-to-face meeting, which proves quite impractical in many circumstances when taking distance and time into account. If one assumes that security is a risk to begin with due to the desire for a secret exchange of data in the first place, the exchange of keys becomes further complicated [5].

Another problem concerns the compromise of a private key. [5] In symmetric-key cryptography, every participant has an identical private key. As the number of participants in a transaction increases, both the risk of compromise and the consequences of such a compromise increase dramatically. Each additional user adds another potential point of weakness that an attacker could take advantage of. If such an attacker succeeds in gaining control of just one of the private keys in this world, every user, whether there are hundreds of users or only a few, is completely compromised.

3. Public-key Encryption

3.1 Summary

Küchlin introduces the foundations of public-key encryption and presents RSA as an early method of transmitting secret messages over insecure channels [5]. The author recognizes that unauthorized users can attempt to intercept messages, and devises this public-key method for ensuring that such users will not be able to interpret the contents of the message [5]. The author's public-key method consists of separate encryption and decryption keys, with users only being able to decrypt an encrypted message if they have the appropriate decryption key [5]. Users will exchange public keys; this transaction does not need to be done in a secure manner because the release of public keys does not threaten the security of any private information. After this swap, someone who wishes to

send private information to another user will encrypt the data with the intended recipient's public key and then pass along the encrypted message. The recipient, who will keep his or her private key secure under any circumstance, can use the private key to decrypt the encoded message. Küchlin introduces separate algorithms for generating encryption and decryption keys as well as an algorithm for combinations of encryption and decryption keys [5].

3.2 Strengths

The asymmetric nature of public-key cryptography allows it a sizable advantage over symmetric-key algorithms. The unique private and public keys provided to each user allow them to conduct secure exchanges of information without first needing to devise some way to secretly swap keys. This glaring weakness of secret-key cryptography becomes a crucial strength of public-key encryption [5].

3.3 Weaknesses

Keys in public-key cryptography, due to their unique nature, are more computationally costly than their counterparts in secret-key cryptography. Asymmetric keys must be many times longer than keys in secret-cryptography in order to boast equivalent security [5]. Keys in asymmetric cryptography are also more vulnerable to brute force attacks than in secret-key cryptography. There exist algorithms for public-key cryptography that allow attackers to crack private keys faster than a brute force method would require. The widely used and pioneering RSA algorithm has such an algorithm that leaves it susceptible to attacks in less than brute force time [3]. While generating longer keys in other algorithms will usually prevent a brute force attack from succeeding in any meaningful length of time, these computations become more computationally intensive. These longer keys can still vary in effectiveness depending on the computing power available to an attacker.

Public-key cryptography also has vulnerabilities to attacks such as the man in the middle attack [3]. In this situation, a malicious third party intercepts a public key on its way to one of the parties involved. The third party can then instead pass along his or her own public key with a message claiming to be from the original sender. An attacker can use this process at every step of an exchange in order to successfully impersonate each member of the conversation without any other parties having knowledge of this deception. [3]

3.4 Proposals

Herzberg, et al. realize the problems presented by the necessity of keeping a private key used in public-key cryptography secret for a long time, and present proactive public key systems that requires more successful hacker attacks in a shorter period of time in order to obtain the private key. Their method builds on threshold cryptography, which they introduce as a method where many users receive parts of the key in order to

protect against any single failure point, but they understand that attackers will still have plenty of time to break the system in certain cases. The paper presents a proactive system that updates the shares periodically in such a way that they are renewed but their shared secret does not change. This robust system meaningfully protects the key, but it does so by transferring the emphasis on security to external hosts. It assumes the security on the servers in which the shares are stored is sufficient, which in a large scale operation is usually sufficient. For more typical users, however, having robust security in several places is a more difficult requirement to meet [4].

4. Digital Signatures

Digital signatures act as a verifiable seal or signature to confirm the authenticity of the sender and the integrity of the message. Users who wish to verify their identity when sending a protected message can encrypt the information with their private key. The recipient can then decrypt the message with the sender's public key in order to confirm the sender's identity and the integrity of the message [1].

4.1 Strengths

Digitally signing a message protects the message in that even if someone intercepted the message before it reached the intended destination and modified it, the digital seal would be broken and the recipient would have this realization after attempting to verify the seal with the sender's public key. The digital signature proves the identity of the sender because only the true sender would have been able to sign the message with his or her private key, except in the event of a compromise. [1]

4.2 Weaknesses

The most serious problems with digital signatures stem from their lack of inherent time stamping. If an unauthorized entity gains access to someone's private key, he or she could send an array of fake messages and sign them with someone else's private key, successfully posing as that other person. The individual whose private key was stolen is unable to repudiate the false messages without having to start over and generate a new private key. To complicate matters, it is impossible to intrinsically separate the fake messages from real ones sent before the compromise because of the absence of time stamping in digital signatures. [1]

4.3 Proposals

Booth examines a proposed encryption method using double encryption, in which a user who wishes to send an encrypted message to another user will encrypt the message with his or her own private key and with the user's public key. The receiver will then decrypt the message using his or her own private key and the sender's public key. This article realizes the problem posed by compromised keys, as either user's private key

falling into the wrong hands can lead to disaster. It proposes a central authentication server, which will receive encrypted messages directly from users, verify that the message has been signed with the sender's current private key, and then attach the receiver's public key and forward the message to its intended destination. This authentication method supercedes the need for an authentication server with a network clock or an archive of compromised keys because as long as it receives notice of all compromises, previous messages will have already been validated [1].

5. Certificate Authorities

Certificate authorities act as trusted third parties that verify the identity of the sender of an encrypted message and issue digital certificates as evidence of authorization. These digital certificates contain the public key of the sender, which is then passed along to the intended recipient.

5.1 Strengths

The issuing of digital certificates allows certificate authorities to play an important role in preventing man in the middle attacks. [3] Certificate authorities have been implemented in the online environment in protocols such as Secure Socket Layer (SSL) and its successor Transport Layer Security (TLS), which have been improved security in web browsing, email, and other methods of data exchange.

5.2 Weaknesses

While certificate authorities aid greatly in the realm of security, they also serve as another potential point of attack. Certificate authorities can be vulnerable to attackers in certain scenarios, and when compromised, can be forced to issue false certificates. Man in the middle attackers who succeed in compromising a certificate authority can use these false certificates to discreetly impersonate each member of the information exchange. Users who are deceived will be even less likely to suspect anything than in a normal man in the middle attack, given the assumed security of the certificate authority. [3]

5.3 Proposals

Halevi and Krawczyk explore an asymmetric case where an authentication server holds private keys while users use only passwords as authentication. They define a password-based authentication protocol where the server uses its own public key to authenticate the user's password, rather than using the user's password as a key to the cryptographic function, which would be a vulnerable and ineffective solution. They also look at a similar approach that uses mutual authentication in which the server possesses both public and private keys, and the user and server authenticate each other. The authors prove that while such systems could still be susceptible to a typical man in the middle approach, where a hacker intercepts messages and replaces them with his own in order to gain an advantage, or other online hacker scenarios, hackers would gain no

added advantage from using an offline password guessing approach that uses computational power to find meaningful patterns, which can be a more effective approach than online attacks against some other security methods [3].

Boyarsky analyzes Halevi and Krawczyk's paper and discovers that their proposal of an asymmetric user-server relationship using server keys and a public password can become insecure when multiple users are introduced to the user-server scenario, with impersonation becoming a real possibility [2]. Boyarsky examines the break in the previous solution, identifying the break as an attacker possibly simulating successful user logins and using this ability to learn the secret password [2]. Boyarsky proposes using the server's public key for signing a user's session key. This system would employ one-time keys, with both the server and user choosing fresh private and public keys for the exchange, which is performed on the user's password [2]. This approach expands on Halevi and Krawczyk's method, satisfying the weakness through the additional key exchange.

6. Future Work

Future work could be done on Herzberg et al.'s idea concerning proactive public key systems. Research in this area could be performed to assess the idea's practicality. Their research holds numerous optimistic positions that would correct the problems surrounding a third party relationship. Positive results in this analysis could lead to a possible implementation of their research in the future.

7. Conclusions

Public-key cryptography has evolved from early models such as Küchlin's to more sophisticated systems that have provided the privacy and data security that we need in the modern world. Secret-key cryptography lags behind asymmetric cryptography. Combinations of the two can be implemented for improved security but secret-key cryptography by itself proves insecure against man in the middle attacks. Asymmetric cryptography has been the foundation for secure data exchange over networks and while it still has its shortcomings, new ideas still come forth as the field continues to evolve.

References

[1] Kellogg S. Booth, "Authentication of signatures using public key encryption," Communications of the ACM, November 1981, pp. 772-774

[2] Maurizio Kliban Boyarsky, "Public-key cryptography and password protocols: the multi-user case," Proceedings of the 6th ACM conference on Computer and communications security CCS '99, November 1999, pp. 63-72.

[3] Shai Halevi and Hugo Krawczyk, "Public-key cryptography and password protocols," ACM Transactions on Information and System Security, August 1999, pp. 230-268.

[4] Amir Herzberg, Markus Jakobsson, Stanisław Jarecki, Hugo Krawczyk, Moti Yung, "Proactive public keys and signature systems," Conference on Computer and Communications Security, 1997, pp. 100-110

[5] W. Küchlin, "Public key encryption," ACM SIGSAM Bulletin, August 1987, pp. 69-73.

Face Recognition

Jonathan Bruno
Department of Computing Sciences
Villanova University, Villanova, PA 19085
CSC 3990 – Computing Research Topics
jonathan.bruno@villanova.edu

Abstract

Biometrics is the automated identification of a person based on physical traits. One biometric which has received considerable attention in recent years is face recognition. Face recognition is considered to be one of the most challenging biometrics because it depends on variations in image quality, orientation, and the subject's appearance. This paper discusses current implementations using 2D or 3D based recognition. 2D recognition achieves generally impressive results. However, accuracy decreases drastically when the images being compared have significant variations. Currently, there is much research being done in the area of 3D recognition which hopes to improve upon the inherent limitations of 2D recognition.

1. Introduction

Face recognition is an attractive biometric for use in security applications. Face recognition is non-intrusive, it can be performed without the subject's knowing. This has become particularly important in modern times because demand for enhanced security is in public interest.

2. Facial Recognition Approaches

2.1 Eigenface-based Recognition

2D face recognition using eigenfaces is one of the oldest types of face recognition. Turk and Pentland published the groundbreaking "Face Recognition Using Eigenfaces" in 1991 [1]. The method works by analyzing face images and computing eigenfaces, which are faces composed of eigenvectors. Results obtained by comparing eigenfaces are used to identify the presence of a face and its identity.

There is a five step process involved in the system developed by Turk and Pentland. First, the system needs to be initialized by feeding it a training set of face images. These are used to define the face space which is a set of images that are face-like. Next, when a face is encountered, the system calculates an eigenface for it. By comparing it with known faces and using some statistical analysis, it can be determined whether the image presented is a face at all. Then, if an image is determined to be a face, the system will determine whether it knows the identity of the face or not. The optional final step concerns frequently encountered, unknown faces, .which the system can learn to recognize.

The eigenface technique is simple, efficient, and yields generally good results in controlled circumstances [1]. The system was even tested to track faces on film. However, there are some limitations of eigenfaces. There is limited robustness to changes in lighting, angle, and distance [6]. Also, it has been shown that 2D recognition

systems do not capture the actual size of the face, which is a fundamental problem [4]. These limits affect the technique's application with security cameras because frontal shots and consistent lighting cannot be relied upon.

2.2 3D Face Recognition

3D face recognition is expected to be robust to the types of issues that plague 2D systems [4]. 3D systems generate 3D models of faces and compare them. These systems are more accurate because they capture the actual shape of faces. Skin texture analysis can be used in conjunction with face recognition to improve accuracy by 20 to 25 percent [3]. The acquisition of 3D data is one of the main problems for 3D systems.

2.3 How Humans Perform Face Recognition

It is important for researchers to know the results of studies on human face recognition [8]. This information may help them develop ground breaking new methods. After all, rivaling and surpassing the ability of humans is the key goal of computer face recognition research. The key results of a 2006 paper "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About" [8] are as follows:

1. Humans can recognize familiar faces in very low-resolution images.
2. The ability to tolerate degradations increases with familiarity.
3. High-frequency information by itself is insufficient for good face recognition performance.
4. Facial features are processed holistically.
5. Of the different facial features, eyebrows are among the most important for recognition.
6. The important configural relationships appear to be independent across the width and height dimensions.
7. Face-shape appears to be encoded in a slightly caricatured manner.
8. Prolonged face viewing can lead to high level aftereffects, which suggest prototype-based encoding. See Figure 1 for an example

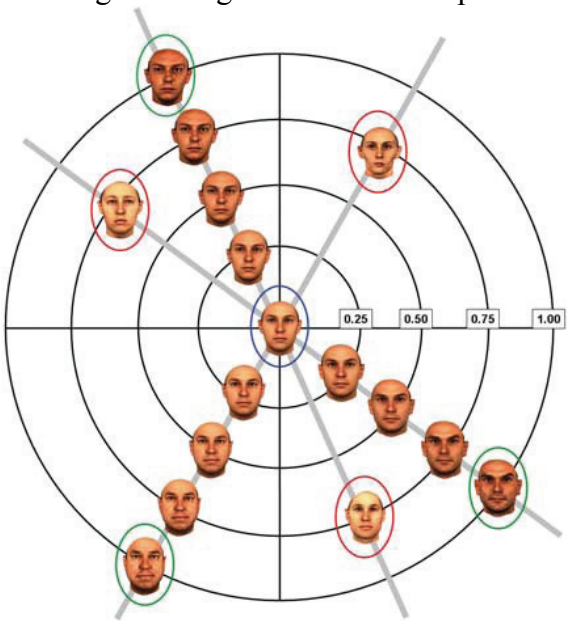


Figure 1. Staring at the faces in the green circles will cause one to misidentify the central face with the faces circled in red [8].

9. Pigmentation cues are at least as important as shape cues.
10. Color cues play a significant role, especially when shape cues are degraded.
11. Contrast polarity inversion dramatically impairs recognition performance, possibly due to compromised ability to use pigmentation cues. See Figure 2.

Figure 2. Photograph during the recording of "We Are the World." Several famous artists are in the picture including Ray Charles, Lionel Ritchie, Stevie Wonder, Michael Jackson, and Billy Joel though they are very difficult to identify [8].

12. Illumination changes influence generalization.
13. View-generalization appears to be mediated by temporal association.
14. Motion of faces appears to facilitate subsequent recognition.
15. The visual system starts with a rudimentary preference for face-like patterns.
16. The visual system progresses from a piecemeal to a holistic strategy over the first several years of life.
17. The human visual system appears to devote specialized neural resources for face perception.
18. Latency of responses to faces in inferotemporal (IT) cortex is about 120 ms, suggesting a largely feed forward computation.
19. Facial identity and expression might be processed by separate systems.

3. Uses of Face Recognition

3.1 Use of Face Recognition

Facial recognition is attractive for law enforcement. It can be used in conjunction with existing surveillance camera infrastructure to hunt for known criminals. Face recognition is covert and non intrusive, opposed to other biometrics such as fingerprints, retina scans, and iris scans [6]. This is especially important in conjunction with the law because faces are considered public. Comprehensive photo databases from mug shots or driver's licenses already exist.

Because of difficulties face recognition has with respect to lighting, angle, and other factors, it is advantageous to attempt to get as high quality images with regard to these factors. Facetraps use strategically placed cameras in order to obtain relatively controlled photographs [6]. Examples are placing cameras facing doorways, at airport

check-ins, or near objects people are likely to stare at (see Figure 3). This type of traps would aid face recognition software by helping to capture a straight frontal image which allow for higher accuracy of the system. Despite their potential benefit, there appears to be very little research done on facetraps.

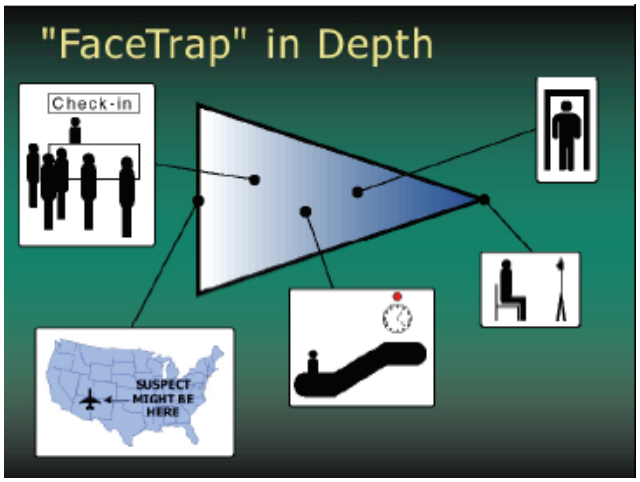


Figure 3. Figure depicts increasingly controlled environments from left to right [6].

Some have questioned the legality of face scanning and have argued that such systems which are used to hunt to criminals in public places are an invasion of privacy. From a legal perspective, in the United States, one does not have a right to privacy for things shown in public [6]. For example; these excerpts from Supreme Court decisions help to establish that face recognition is constitutional. "What a person knowingly exposes to the public. . . is not a subject of Fourth Amendment protection," United States v. Miller, 425 U.S. 435 (1976). "No person can have a reasonable expectation that others will not know the sound of his voice, any more than he can reasonably expect that his face will be a mystery to the world," United States v. Dionisio, 410 U.S. 1 (1973).

Face recognition must be improved further before it becomes a useful tool for law enforcement. It remains to be seen what the right balance is, socially speaking, between maximizing public safety and respecting individual rights.

3.2 Other Uses of Face Recognition

Implementations of face recognition systems include surveillance cameras in Tampa, Florida and Newham, Great Britain [2]. Trials of the systems yielded poor results. The Newham system did not result in a single arrest being made in three years. Logan Airport, in Boston, performed two trials of face recognition systems. The system achieved only 61.7% accuracy [5]. Australian customs recently rolled out its SmartGate system to automate checking faces with passport photos. Google is testing face recognition using a hidden feature in its image searching website [7]. Google purchased computer vision company Neven Vision in 2006 and plans to implement its technology into its Picasa photo software.

4. Future Work

Face images which appear in databases are taken in controlled environments. Current face recognition technology has difficulty comparing faces which vary in angles

or lighting. Recent deployments of face recognition systems have yielded poor results because faces captured in the images vary widely from the database images. One way remedy to this situation is to use facetraps. Facetraps are cameras which are strategically placed to capture high quality images of faces. The goal is to obtain images which are as close as possible to those taken in the controlled environment.

Figure 4. Hidden cameras which look like everyday items will be useful to ensure subjects are unaware of the cameras (electrical box and wall clock with hidden cameras are pictured).

Our proposal will determine the effectiveness of different facetrap setups. Several facetrap scenarios will be tested in a busy, public area. Some set ups which will be tried are placing cameras facing doorways, near clocks, behind check-out counters, and behind one way mirrors. It is imperative that hidden cameras be used so that subjects do not realize they are being watched. The cameras will collect data for two months. Image quality will be judged on angle, lighting, and distance. Facetraps which consistently yield good results will be noted as good candidates for actual implementation. Further work may involve new ideas for facetrap placement or tweaks to previously tested methods.

References

[1] Matthew A. Turk, Alex P. Pentland, "Face Recognition Using Eigenfaces," Proc. IEEE Conference on Computer Vision and Pattern Recognition: 586–591. 1991.

[2] Michael Kraus, "Face the facts: facial recognition technology's troubled past--and troubling future," The Free Library, 2002.

[3] Mark Williams, "Better Face-Recognition Software," Technology Review, May 30, 2007.

[4] Trina D. Russ, Mark W. Koch, Charles Q. Little, "3D Facial Recognition: A Quantitative Analysis," 38th Annual 2004 International Carnahan Conference on Security Technology, 2004.

[5] Ryan Johnson, Kevin Bonsor, "How Facial Recognition Systems Work," How Stuff Works, 2007.

[6] John D. Woodward, Jr., Christopher Horn, Julius Gatune, Aryn Thomas, "Biometrics, A Look at Facial Recognition," RAND, 2003.

[7] "New: Google Image Search Categories," Google Blogoscoped, May 28, 2007.

[8] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell, "Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About," Proceedings of the IEEE, Volume: 94, Issue: 11, 2006.

Interference Reduction in Wireless Networks

Joseph Bruno

Computing Research Topics
Computing Sciences Department
Villanova University
Villanova, PA
joseph.bruno@villanova.edu

November 14, 2007

Abstract

The presence of interference in wireless networks is quickly being focused upon as a problem in today's world. Unfortunately, research is so young that leading researchers have differing ideas regarding the identification of a universal measure of network interference. But some goals have become cemented as necessary, such as the requirement of a low interference network that does not sacrifice other aspects of a network and maintains overall network performance. One proposed methodology of interference reduction is through topology control, which seeks to establish a strong network while still keeping the interference generated at a minimum. In this paper, we discuss existing topology control algorithms and present open problems to spur on active development of interference reducing network solutions.

1. Introduction

With the growing concern of a shrinking world and the advancement of modern technology, the use of wireless networks has spread rapidly over the last few years. From that growth, the threat of an interference-induced loss of quality in wireless networks is becoming more imminent. Interference, caused by multiple transmission signals being sent simultaneously over a shared radio frequency, induces a loss of data being sent over a network. Due to interference, the packets of data sent over wireless networks become lost and never make it to their destination, requiring more time and energy sent to resend the packets until they reach their final destinations. Research is currently under way to attempt to find ways to reduce wireless network interference, and allow for a continued use of wireless networks without worry of a future where such technology will be rendered ineffectual. This research is performed through topology control, a way to structure network placement and signal transmissions in a given area in order to minimize the interference in that area. Before reducing interference in an area, however, interference must be measured so that we can tell how much of an improvement is made using a particular method of reduction.

2. Measurement of Interference

Interference is typically measured in two ways, traffic-based and traffic-independent. Most uses of topology control define interference as "traffic-based interference," or measured in the airspace of a network through the number of overlapping signals of the same frequency and their effect on surrounding nodes. Since traffic-based interference is application-dependent, "traffic-independent" interference measures have been developed. According to Pascal von Richenbach et. al [2], the only

way to accurately define and view interference is looking at topologies with traffic-independent measurement in mind. Such models depend on the network topology only.

The most common (and simpler) topology is the Unit Disk Graph (UDG), with wireless senders and receivers as vertices (also called nodes), and edges connecting pairs of nodes that can communicate directly with each other.

One particular interference model is the sender-centric model, which considers interference as an issue at the sender's location. Claiming that such a model does not reflect upon real world interference very well, receiver-centric model was formed, which defines interference as the number of nodes that contain another node (the receiver in this case) within their communication range. The reasoning behind this is that the sender does not actually experience the interference when sending the signal, and that the receiver is subject to a poor quality reception due to interference.

In light of debates regarding a universally agreed upon system of interference measurement and definition, algorithms addressing interference reduction state the measure of interference that is being used. This freedom from a particular measurement constraint, however, allows more potential solutions to be developed.

3. Methods for Reducing Interference

Currently, the field of wireless interference reduction is young, and research scant. This lack of much background work has allowed for a multitude of perspectives to be taken towards interference reduction. A common flaw according to Rickenbach [2] is that the vast majority of interference-reducing algorithms concentrate on simply reducing the number of links in a network, believing that with fewer signals present the interference will "sort itself out." This idea, however, is not the most effective at generating a lower interference. Similarly, Burkhart [1] [3] claims that the maximum degree in a network serves merely as a lower bound for the network's interference. From this, it can be concluded that algorithms that concentrate on reducing interference through methods other than simply reducing the degree of nodes in a network should be sought after.

While some algorithms do attempt to follow these declarations, the vast majority still adhere to the older, more basic target of link reduction. The end goal of earlier algorithms centering on link reduction is to find a balance between network interference present and the time and power used to reduce interference further. While there are ways to reduce interference significantly, these may come at the cost of draining large amounts of power or overly long time taken to generate the topology. While these methods may generate a significantly interference-light network area, the means taken to implement them are too costly to justify the results.

Figure 1 demonstrates an example network where node v connects to node u directly, which would cause interference at all nodes that lie within the disk centered at v of radius uv. Figure 2, on the other hand, shows this same situation, only with node v using intermediate nodes to access node u while generating significantly less interference. Figure 3 shows an example that uses too many links between two nearby nodes to justify the resulting reduced interference. With this balance as a goal in mind, several different topologies have been developed.

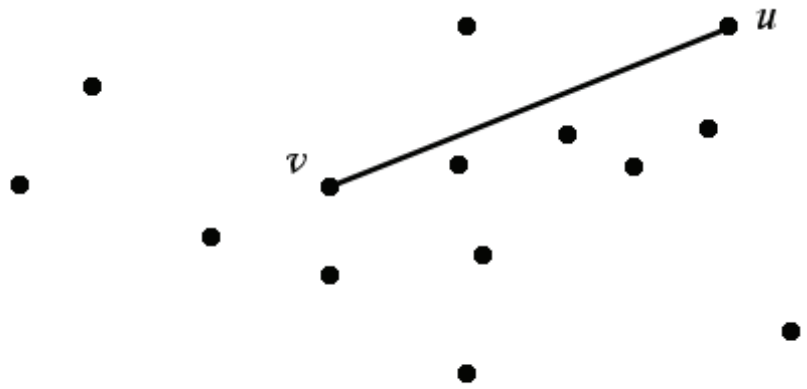


Fig. 1. A high-interference topology with directly connected nodes.

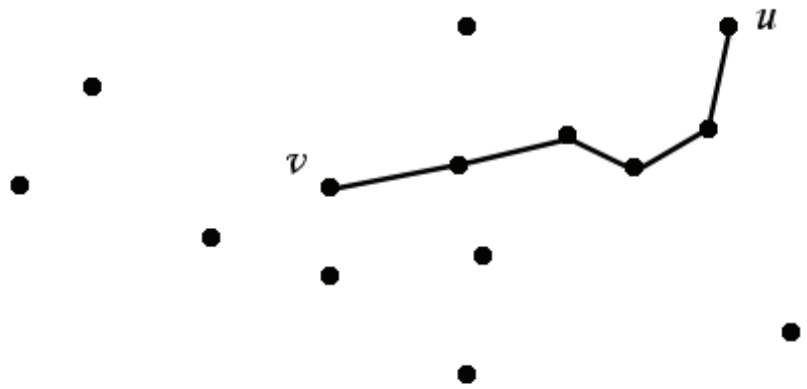


Fig. 2. A low-interference topology network utilizing intermediate shorter links in place of a long link.

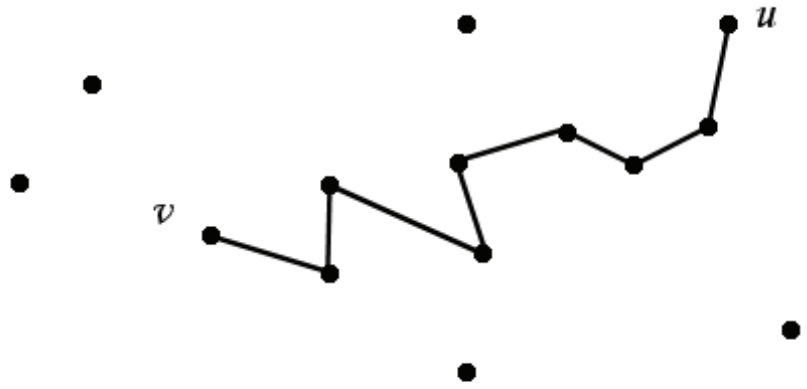


Fig. 3. A topology using too many intermediary nodes to justify a reduced interference.

4 Algorithms

One question that is still very much in theoretical stages is what topology control algorithm to implement to obtain the optimal interference in a given network, while still preserving its connectivity. Connectivity is defined as the ability of one wireless node to communicate with any other node. All current algorithms require wireless nodes to extend their transmission range to a specific length, in efforts to establish connections between nodes without the wireless signal going any further than necessary and causing interference at other nodes.

4.1 Greedy Algorithms

A very simple topology control algorithm would require nodes to perform a search of their surrounding area and establishing a connection with the closest one. While this "greedy algorithm" successfully reduces interference, Wattenhoffer [4] shows that it does not generate optimal interference for a network. An example in which such an algorithm performs relatively poorly is an exponential node chain, which is covered in Section 4.2.

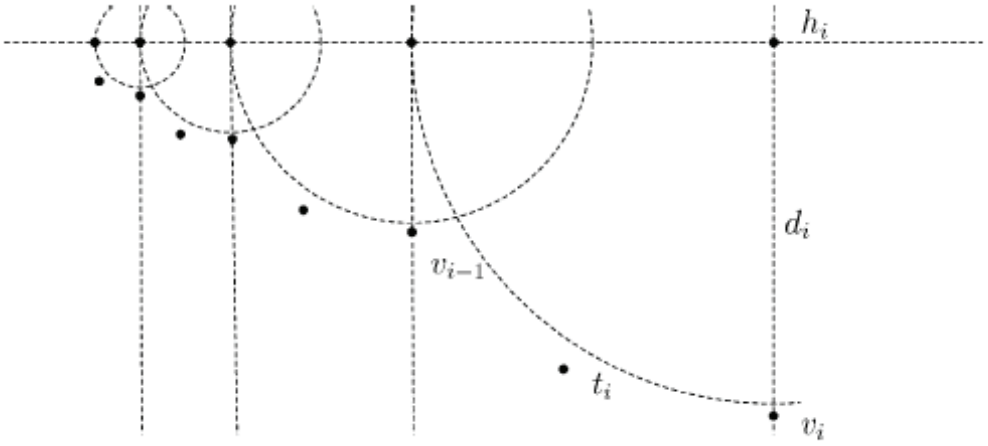


Fig. 4. A network where the greedy algorithm does not yield optimal interference [1].

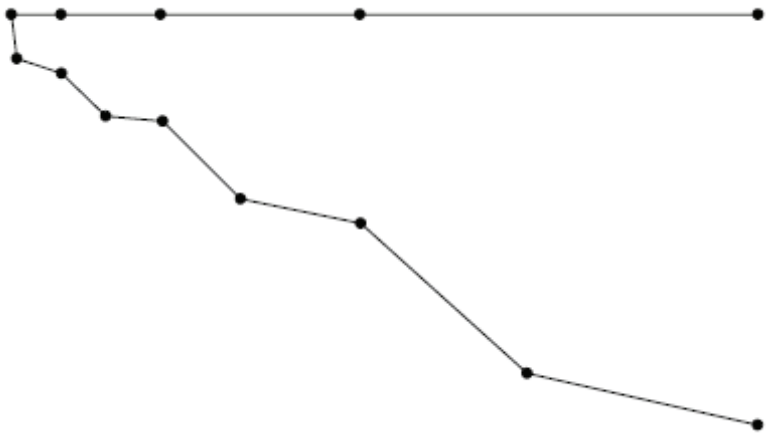


Fig. 5. The closest-neighbor method for this two exponential node chain example produces interference O(n) [1].

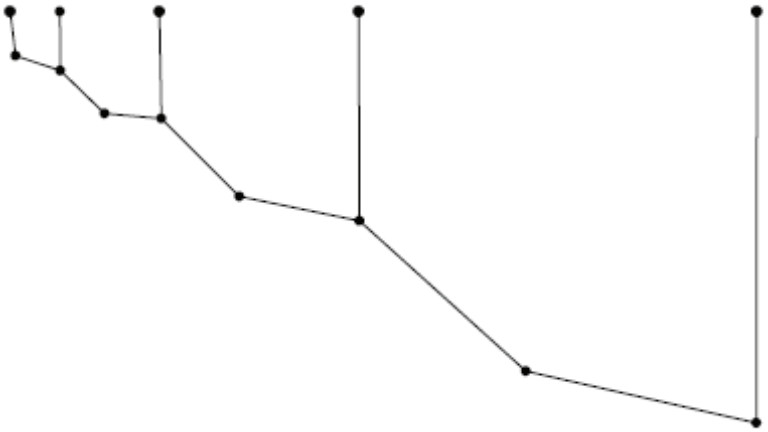


Fig. 6. The same node set as in Fig. 5, connected in optimal interference topology [1].

Wu and Liao [5] developed algorithms that improve upon this greedy algorithm. The work of Wattenhoffer, Wu, and Liao demonstrates that the low-interference topology control problem is indeed very difficult, and in light of this, algorithms for special network topologies, such as the exponential node chains, have been developed.

4.2 Exponential Node Chains

Topology algorithms for reducing receiver-centric interference are those designed around the "exponential node chain," a topology of nodes that are placed in a straight line with the distance between each pair of nodes growing exponentially longer. A baseline algorithm used in such topologies connects each node with its closest neighbor (Fig. 4). Due to the increasing amount of overlapping signals, this algorithm results in O(k) interference at the *k*th node from the left, or O(n) interference overall. Another network example consisting of two exponential node chains, shows that connecting each node to its closest neighbor (see Fig. 5) results in O(n) interference, however the optimal topology (as depicted in Figure 6) has constant interference. Rickenbach [6] improves

upon this baseline algorithm by developing a clever algorithm that results in O(√n) interference, which is optimal for the exponential node chain (displayed as in Fig. 7). The degree of each node in this topology is O(√n).

Fig. 7. Pascal von Rickenbach's logarithmic exponential node chain topology [6]. Here, each hollow node is a hub and is connected to every other node, as well as their nearby hubs.

4.3 The Highway Model

A generalization of the exponential node chain is the Highway Model, a linear topology where the nodes are not necessarily arranged so that the distance between each pair is exponentially longer than the last. Although this model topology seems rather simple, approximation solutions for this model are fairly complex, and optimal solutions are yet to be developed. Rickenbach [6] has proposed an approximation solution that employs a decomposition of a "highway" into multiple exponential node chains and uses the optimal algorithms designed for such chains., The algorithm employs virtual nodes that will later be adjusted so the signals to the virtual nodes can translate to the true nodes present. Another solution proposed by Rickenbach [2] involves the highway to be split into segments of equal length, with certain nodes designated as hubs. These hubs will connect to each other in a linear fashion, as other nodes connect to the nearest hubs, as in Figure 8. If the largest number of exponential node chains in a segment is *k*, the resulting topology has $O(\sqrt{k^3})$ interference.

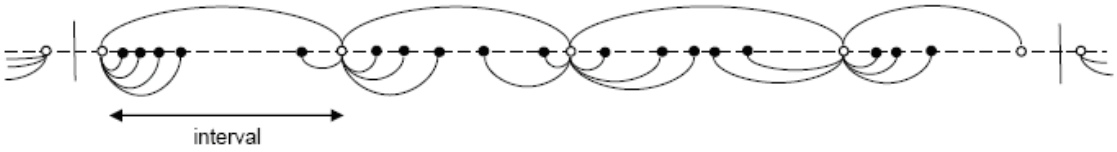


Fig. 8. Rickenbach's second proposed solution to the Highway Model [2]. Each hollow node is a hub, and each node is connected to its nearest hub.

5. Proposed Work

We propose to expand upon previous algorithms designed for exponential node chains so they may be usable in a two dimensional topology. The difficulty of forming such an algorithm is the reason why little to no research on making more universally applicable exponential node chain algorithms has been conducted.

Plans for the development of an exponential node chain algorithm so that it may function in two dimensions utilize sectioning off the surrounding area of an arbitrarily chosen root node of a topology. Once the root is designated, a cone-shaped signal will be

sent from the root node, incrementing itself around the node so that no two sections overlap. At most, there will be three hundred and sixty sections around the root node, assuming each section is set for one degree of width. The signals will perform exponential node chain construction within the current section, with the root node of the network acting as the chain's root, and then perform an interference analysis of the resulting topology.

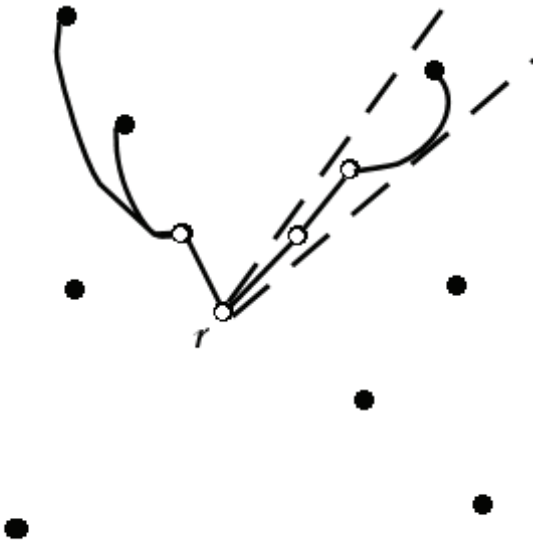


Fig. 9. Root node *r* sends constructs an exponential node chain of the topology within the current section denoted by the dotted lines. Following Figs. 7 and 8, each hollow node is a hub, and each node is connected to its nearest hub.

The work on such an algorithm is projected to take up to two months of time. The first week will be spent concluding all current research on wireless network interference reduction. The following two weeks will be used to determine topologies that can be simulated for this algorithm with programs such as the free QualNet [7]. The following three weeks will be spent developing the algorithm, as well as testing the algorithm's implementation and performance. The final week will be centered around refining the algorithm in order to produce the most efficient and quickest result, while still ensuring the best possible interference in a wireless network.

6. Conclusion

This paper reviewed some of the currently existing methods designed to reduce the interference present in wireless networks. Through an overview of both current and past research, the groundwork has been laid for the formation of a new algorithm designed to improve upon the current Exponential Node Chain algorithms. The future of interference reduction has many open questions requiring answers, but appears that there is promise due to the growing concern of interference's existence in a technologically growing global society.

Sources

[1] M. Burkhart et. al, "Analysis of Interference in Ad-Hoc Networks," diploma thesis, Distributed Computing Group, Institute for Pervasive Computing, 2003, pp.7-42.

[2] P. von Rickenbach et. al, "A Robust Interference Model for Wireless Ad-Hoc Networks," 5th International Workshop on Algorithms for Wireless, Mobile, Ad Hoc and Sensor Networks (WMAN), April 2005, pp. 3-8.

[3] M. Burkhart et. al, "Does Topology Control Reduce Interference?", Int'l Symp. Mobile Ad Hoc Networking & Computing, 2004, pp. 9-19.

[4] R. Wattenhofer, "Algorithms for ad hoc and sensor networks," Elsevier Journal on Computer Communications, Volume 28, Issue 13, August 2005, pp. 2-6.

[5] K. Wu, W. Liao, "Interference-Efficient Topology Control in Wireless Ad Hoc Networks," Consumer Communications and Networking Conference, 8, Jan. 2006, pp. 411-415.

[6] P. von Rickenbach et. al, "Interference and Topology Control in Ad-Hoc Networks," master's thesis, Distributed Computing Group, Institute for Pervasive Computing, 2004, pp. 33-48.

[7] http://www.scalable-networks.com/products/developer/new_in_40.php

Interference Reduction in Wireless Networks Using Graph Coloring Methods

Andrew Chickadel
Computing Research
Department of Computer Science
Villanova University
Villanova, PA 19085
Andrew.Chickadel@villanova.edu

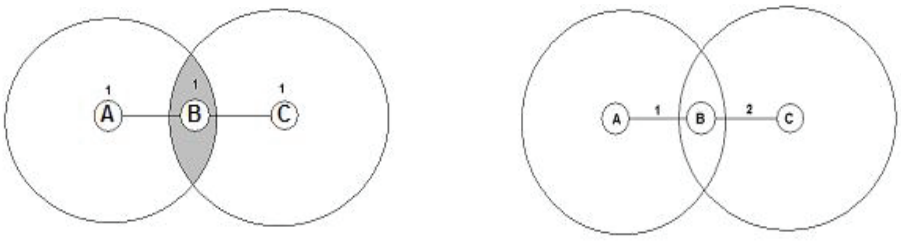
Abstract

The problem of reducing interference in wireless networks is becoming increasingly important with the continuous deployment of larger and more sophisticated wireless networks. In this report we analyze the interference problem from a graph theoretical viewpoint. The interference reduction problem is modeled as a graph coloring problem. However, additional constraints to graph coloring scenarios that account for various networking conditions result in additional complexity to standard graph coloring. This paper discusses a variety of algorithmic solutions for specific network topologies and makes recommendations for implementation.

1. Introduction

Signal interference is a major drawback of wireless networks. The aim of reducing interference is to prevent adjacent or connected nodes, which are linked by radio signals, from receiving and transmitting signals which conflict or blend together. Thus, interference occurs when conflicting transmissions over one radio frequency are received by one or more nodes in a wireless network. This inhibits the ability of the receiver to decipher incoming signals. This concept is illustrated in Figure 1a, which shows a typical situation in which the broadcast areas of nodes A and C overlap in the vicinity (shaded in gray) of node B, causing B to receive a garbled signal composed of the signals from A and C. In such situations, it is difficult for B not only to decipher simultaneous signals, but also to reliably determine the source of the signal. Applying the graph coloring problem to reducing interference in wireless networks is practical compared to investing large amounts of money in radio transmitter technology or simply hard-wiring a temporary fix. The problem of reducing interference in arbitrary networks turns out to be very difficult, and for this reason, simpler network layouts have been investigated such as, multi-hop wireless mesh network layouts [1], triangular lattice topologies [9], unit disk graphs [3], hexagonal topologies [8], and other more general topologies [4]. Other key facets of the interference problem in wireless networks specify whether a proposed solution is contrived in a distributed or centralized setting, whether nodes in a given solution are self-aware of their location or whether this assumption is not necessary, and

whether or not minimum separation distance between nodes needs to be factored into algorithmic solutions.



(a) (b)

Figure 1. Interference resolved through channel assignment. (a) Interference occurs at B. (b) No interference.

Numerous methods for reducing interference exist, such as topology control [10], power control [7], and channel assignment [1, 2, 3, 5, 6, 8, 9]. This paper will focus exclusively on the latter method, which seeks to assign channels of different frequencies to interfering nodes or edges.

In Figure 1a, for instance, simultaneous transmissions of A and C result in interference at B. This problem is resolved in Figure 1b by having nodes A and C transmit over frequencies 1 and 2 respectively, equipping node B with two radios that can transmit and receive over frequencies 1 and 2. Signals from A and C can be demultiplexed, (that is, components of different frequencies can be extracted from one signal) at node B because of differing frequencies, and node B can clearly determine if node A transmits across frequency 1 or if node C transmits across frequency 2. Therefore the intersection node A and node C's broadcast areas no longer results in interference.

Through careful assignment of communication channels to nodes in a network, interference could be greatly reduced. It is important to note, however, that the number of radio frequencies is finite, and therefore, the problem of minimizing the number of channels allocated to a specific network is worthy of thorough investigation as well. In some instances, channel overlap is necessary if the number of assigned channels for a network is inadequate to connect all nodes [2].

The rest of the paper is organized as follows. In Section 2, we define the interference problem as a graph coloring problem and discuss two coloring approaches, vertex and edge coloring. Open problems and future work are discussed in Section 3. Finally a summary of existing results and conclusions are presented in Section 4.

2. Vertex vs. Edge Coloring

Interference in wireless networks is commonly represented as an interference graph in which edges represent potential interference between the endpoint nodes. The problem

of reducing collisions and signal interference is modeled as a coloring problem on the interference graph. Nodes of different colors in the graph will be assigned separate channel of radio frequency. Efficient coloring algorithms will lead to an effective channel selection method that lowers the wireless interference.

The graph coloring problem's application is relevant because preventing vertices from connecting via radio frequency with other (conflicting) vertices is the quintessential task in reducing interference. Graph coloring algorithms are also reliable in that they are mathematically provable. Channel assignment, however, adds algorithmic complexity to standalone graph coloring problems, according to Khanna and Kumaran [8]. Channel assignment is indeed what ties the graph coloring problem together with the problem of reducing interference in wireless networks.

General graph coloring problems exist in two distinct varieties, edge and vertex coloring. The main condition for edge coloring problems is that no vertex in a graph can have more than one outgoing edge of a particular color. Edge coloring problems are separated into two classes, one corresponding to bipartite graphs and one corresponding to all other graph scenarios. Vertex graph coloring problems are not separated into classes, but merely require one condition that no vertex in a graph can be colored the same color as an adjacent vertex. Figure 2 below shows an example of vertex coloring using the minimum number of colors possible. Typically, all algorithms implementing the graph coloring problem strive to use the fewest colors possible.

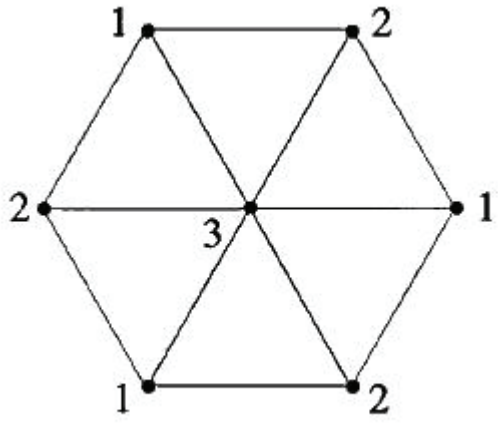


Figure 2 [Extracted from 9]. A sample 3-coloring.

There is ongoing research explicitly investigating vertex-coloring and edge-coloring graph methods that address the channel assignment problem for wireless networks [1, 2, 3, 5, 6, 8, 9]. The order in which colors or channels are assigned and to what nodes differs greatly from one solution to another. The creation of such an order is also quite varied among current proposed solutions. Weighted coloring is one method that deals

exclusively with assigning channels based on need to alleviate interference within the network.

2.1 Weighted Coloring

One variation of the graph coloring problem involves assigning weights to the interference graph. A weighted coloring implementation addresses interference chiefly in areas of greatest need. Once these problem areas are discovered, channel re-assignment can alleviate signal collision. McDiarmid and Reed [9] assign weights based on bandwidth demands, whereas Arbaugh et al. [2] assign weights that indicate the degree of channel interference between two nodes. McDiarmid and Reed [9] reduce interference by assigning to each node a number of colors equal to its weighted bandwidth demand value. Arbaugh et al. [2] devise algorithms, which they call Hminmax and Hsum, which greedily choose the frequency at each node which will locally at that node result in the greatest reduction in interference. McDiarmid and Reed [9] state that there is future work in determining an improved ratio for large demands. Arbaugh et al. [2] state, however, that for their algorithms Hminmax and Hsum, wireless-b and wireless-g environments were not studied in a mixed setting, and this remains an open issue.

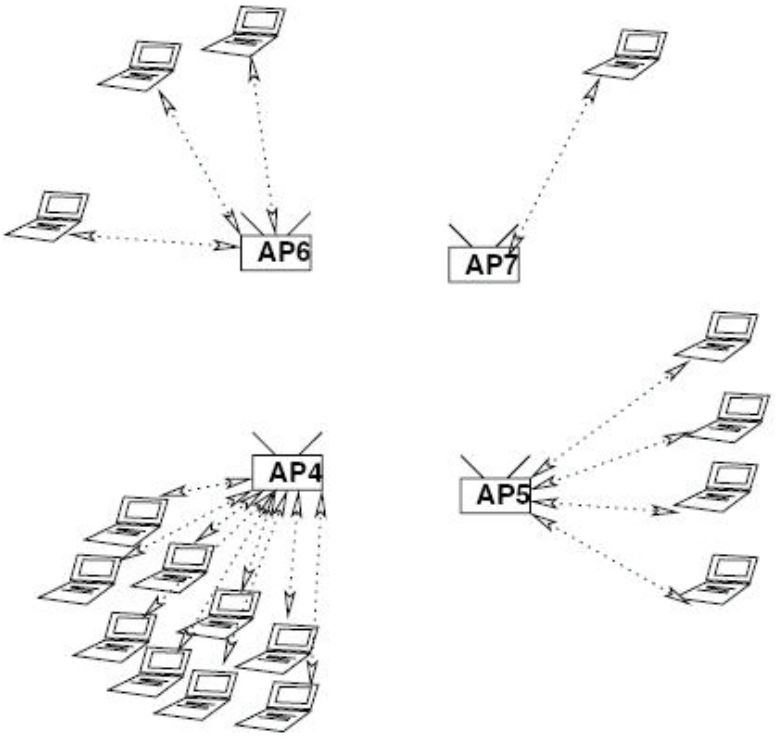


Figure 3. All Access Points (AP) are in interference range of each other [2].

Figure 3 depicts a network where there are only three non-overlapping channels for four wireless access points labeled AP4, AP5, AP6, and AP7. Each access point can

broadcast far enough to reach each wireless device attached to it. In this instance, two access points will need to share a communication channel [2].

A greedy weighted algorithm would assign AP4 and AP5 their own channel to broadcast over, because they have more devices connected than AP6 and AP7 do. AP6 and AP7 therefore will have to share one channel and endure the possibility of interference to spare the majority from the threat of interference. Provided that AP6 and AP7 are sufficiently far from each other, sharing one communication channel may result in no interference. This idea of channel reuse has been explored by several researchers aiming to reduce not only interference, but the number of communication channels as well.

2.2 Channel Re-use

Channel re-use involves imposing a minimum distance between two nodes operating at the same radio frequency. Re-using channels not only curtails interference but also improves overall efficiency in terms of channel use. Bertossi et al. [5] define the channel assignment problem with separation (CAPS) in order to incorporate such a re-use distance in channel assignment, resulting in a more efficient use of the given set of radio frequencies. An algorithm is presented by which a node, knowing its relative position in the network, can compute its channel assignment, with a specified re-use distance, in constant time for all network graph types except for binary trees, which require logarithmic time [5]. Bertossi et al. [5] however, identify an open issue exists in determining an optimal solution for general graphs and arbitrary channel re-use distance. An alternative for computing minimum distance would be to determine the minimum and maximum number of channels to use in a network.

2.3 Radio Frequency (RF) Spectrum

Investigating strict upper and lower bounds for the number of channels used in a given wireless network is of importance to the interference reduction problem. Khanna and Kumaran [8] define what they call the *wireless spectrum estimation problem*, where a node is assigned the smallest number of frequencies over which to broadcast and receive, such that interfering nodes do not share the same frequency. Bertossi et al. [5] later explore channel re-use governed by a minimum re-use distance rather than static assignment as delineated by upper and lower bounds (see Section 3.2). Additionally, the ability to compute minimum distance based on relative position is not always a necessity.

2.4 Location-Oblivious

Location-oblivious networks do not rely on each node knowing its relative geometric position. According to Barbeau et al. [3], location-aware networks, on the other hand, might require a Global Positioning System (GPS) in order to calculate nodes' relative positions. Therefore, location-oblivious networks are preferable. Barbeau et al. [3] introduce a distributed, location-oblivious unit disk graph coloring algorithm unlike Bertossi et al. [5] whose algorithms require nodes to be self-aware of their location.

However, the algorithm presented by Barbeau et al. did not out-perform an algorithm that executes arbitrary coloring of a network graph [3]. Similarly to location-oblivious algorithms, dynamic channel assignment [2] does not require a node to know its relative position.

2.5 Dynamic Channel Assignment

Dynamic channel assignment can reduce interference in networks where the topology changes often and dramatically. Dynamic channel assignment involves continuously monitoring interference and re-assigning channels appropriately. Arbaugh et al. [2] employ a form of dynamic assignment in their multi-step, greedy approach algorithms Hminmax and Hsum that re-assign channels only to nodes experiencing the greatest interference (mentioned in Section 3.1). Almeroth et al. [1] introduce dynamic assignment in their algorithm, Breadth First Search - Channel Assignment (BFS-CA). BFS-CA utilizes breadth-first search to determine nodes of greatest connectivity which typically are subject to the greatest interference. Channels are re-assigned on-the-fly, and connectivity is maintained through a secure "default channel" [1]. Figure 4[1] below illustrates why a default channel is vital to the structure of the network.

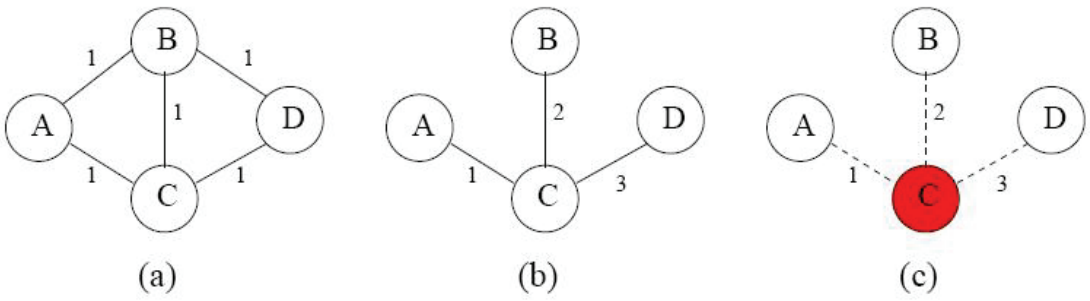


Figure 4. Network topology with varying channel assignments [1].

Figure 4(a) depicts a sample network topology where all nodes broadcast over one frequency. Suppose node C is selected for channel re-assignment via the addition of two new radios that broadcast and receive over channels 2 and 3. Figure 4(b) shows the result of this re-assignment. Through this process of eliminating the threat of interference, nodes A, B, and D lose direct connection to each other and instead must communicate using one-hop over node C. However, further complications arise with this topology change. Figure 4(c) shows the event of node C failing. The dotted lines represent broken connections. The consequence of this is that nodes A, B and D will not be able to communicate, even though they are within communication range of one another. For all changes in topology, node failure is a real threat in breaking communication. With a default channel in place, however, a severed connection due to node failure can be overcome with minimal downtime [1]. In addition to how BFS-CA [1] will perform in very dense topologies, addressing just how long the downtime due to node failure will

last is an open issue, and, in fact, graph coloring algorithms in general are works-in-progress because of their NP-hard and NP-complete algorithmic complexities.

3. Proposed Future Work

For future work, we propose to analyze the performance of the algorithms BFS-CA [1], Hminmax, and Hsum [2] on Unit Disk Graphs (UDG), both empirically (through implementation and experimentation) and analytically. UDGs are much simpler network models compared to typical mesh models, in the sense that only nodes that are no farther apart than the unit distance (after appropriate scaling of the maximum transmission range) can directly communicate to one another and therefore are connected by a link in the network graph.

We propose to take one step further and study the same algorithms on Quasi Unit Disk Graphs (qUDG), which are more realistic than UDGs, yet simpler than general networks. qUDGs take an input parameter alpha, with $0 < \text{alpha} <= 1$, and guarantee links between nodes that have a distance not greater than alpha; nodes at distance greater than 1 are not connected by a link, and nodes at distance in the range (alpha, 1] may or may not be connected. This model attempts to account for errors in transmission, fading signal strength, and other imperfections.

There is much potential for greater interference reduction on UDGs and qUDGs than on general networks. In order to demonstrate such results, we plan to simulate BFS-CA, Hminmax, and Hsum in a wireless mesh network environment using Scalable Network Technologies' QualNet [11]. QualNet is a free network simulator that allows users to easily create, manipulate, and simulate any kind of network model through an intuitive graphical interface. Again, using QualNet, we will then implement BFS-CA, Hminmax, and Hsum and execute these algorithms on UDG graphs with varying number of nodes, say between 100 and 1000.

We will collect statistics such as maximum and average interference, retransmissions, channel reassignment, and propagation delays among others. We expect to see significant improvements in interference reduction by simulating BFS-CA, Hminmax, and Hsum in UDG and qUDG environments compared to simulation in various wireless mesh network environments. All results will be thoroughly documented and compiled onto a web page yet to be created.

4. Conclusion

The dynamic algorithm, BFS-CA [1], is one of the best algorithms for use in today's ever-changing wireless network topologies. It is the most implementation-ready compared to other graph coloring algorithms. BFS-CA was also shown to have a significant improvement over static assignment of channels [1]. The weighted Hminmax and Hsum [2] algorithms, despite resorting to greedy implementations, have achieved over a 40% average reduction in interference over one "state-of-the-art" method [2]. The

McDiarmid and Reed bandwidth-based weighted algorithms [9] bring together several novel ideas, however they seem difficult to eventually implement.

5. References

[1] K.C. Almeroth, E. M. Belding, M. M. Buddhikot, K. N. Ramachandran, "Interference-Aware Channel Assignment in Multi-Radio Wireless Mesh Networks," Apr. 2006; http://www.cs.ucsb.edu/~ebelding/txt/infocom06.pdf.

[2] W. Arbaugh, S. Banerjee, A. Mishra, "Weighted Coloring Based Channel Assignment for WLANs," *ACM SIGMOBILE Mobile Computing and Communications Review*, July 2005, vol. 9, no. 3, pp. 19-31.

[3] M. Barbeau, P. Bose, P. Carmi, M. Couture. E. Kranakis, "Location Oblivious Distributed Unit Disk Graph Coloring" School of Computer Science Carleton University Jan 2006 pp. 1-20.

[4] R. Battiti, A. A. Bertossi, M. A. Bonuccelli, "Assigning Codes in Wireless Networks: Bounds and Scaling Properties," *Wireless Networks*, May 1999, pp. 195-209.

[5] A. A. Bertossi, C. M. Pinotti, R. B. Tan, "Efficient use of radio spectrum in wireless networks with channel separation between close stations," *Workshop on Discrete Algorithms and Methods for MOBILE Computing and Communications* and *Proceedings of the 4th International Workshop on Discrete Algorithms and Methods for Mobile Computing and Communications*, 2000, pp. 18-27.

[6] T. Chiueh, K. Gopalan, A. Raniwala, "Centralized Channel Assignment and Routing Algorithms for Multi-Channel Wireless Mesh Networks," *ACM SIGMOBILE Mobile Computing and Communications Review*, Apr 2004, vol. 8 no. 2, pp. 50-65.

[7] D. J. Goodman, N. B. Mandayam, C. U. Saraydar, "Efficient Power Control Via Pricing in Wireless Data Networks," *IEEE Transactions on Communications*, Feb 2002, vol. 50, issue 2, pp. 291-303.

[8] S. Khanna, K. Kumaran, "On Wireless Spectrum Estimation and Generalized Graph Coloring," *INFOCOM '98. Seventeenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, Apr 1998, vol. 3, pp. 1273-1283.

[9] C. McDiarmid, B. Reed, "Channel Assignment and Weighted Coloring," *Networks,* Aug. 2000, vol. 36, no. 2, pp. 114-117.

[10] R. Ramanathan, R. Rosales-Hain, "Topology Control of Multihop Wireless Networks Using Transmit Power Adjustment," *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, March 2000, vol. 2, pp. 404-413.

[11] Scalable Network Technologies, Inc. QualNet Network Simulator. 2007. http://www.scalable-networks.com/

Wireless Distribution Service Systems

Xavier Halloran
Computing Research
Department of Computer Science
Villanova University
Villanova, PA 19085
daniel.halloran@villanova.edu

Abstract:

This paper identifies problems and offers solutions for large-scale wireless distribution systems in low-power environments. Many situations require fast deployment of low cost, long distance wireless access points in areas where little to no electrical power is available, such as public parks, disaster areas, and construction sites. Issues such as solar power and wind power alongside batteries are considered, along with power efficient protocols. Mechanisms to control bandwidth along distributed wireless networks such as Wireless Distribution Service are also analyzed. Various types of cheap and effective hardware and firmware are also discussed. All of these components are brought together to form a possible practical solution that is power friendly, cost effective, and relatively power stingy.

1.1 History

Wireless systems have been in development for decades now. Technology has advanced from basic radio communication presented as a back pack device and truck mounted equipment all the way to the current minimal size of the Razor phones. From our cell phones to our laptops to satellite television, wireless communication stands inseparable from our lives, and without it we would never have the conveniences of today's world.

2 Importance

2.1 Networking

Networking has revolutionized the world. Since the original telegraph transmissions allowed news to travel across the world in seconds, the world began shrinking. Technology has transformed communication, bringing everyone into such close contact that now, there is an imperative need to be attached to the network. Rapid communication with your friends and allies allows for watchful eyes on your adversaries. With up-to-the minute information at our disposal, rapid response and quick changes of strategy enable effective maneuvering in critical situations.

2.2 Modern Demand

The Internet now exists as an essential tool for any business to succeed, as well as for any modern person to cope in day to day life.

Without network connectivity, and the Internet as a whole, most business would not succeed. Credit would fail and be reduced the pre-electricity days of credit, where consumers are taken at a word rather than history as a result of the lack of electronic data available to check credibility.

2.3 Areas of Interest

In many situations, the deployment of computer networks is critical, yet establishing a wired network is unfeasible. In this case, completely wireless networks, or networks with no significant distance wiring, are required. The arrangement and configuration of the many access points and the possible protocols are areas of interest. Directed mid-range wireless signals, on the order of hundreds of feet, create long-range noise, polluting valuable and limited spectrum. Sharing a wireless spectrum between the back-bone signals (the equivalent of the wired part of a typical wireless network) and the users becomes yet another problem. Data routing in this case is simplified, as all of the access points are geographically static, and paths are constant. Power supply is an issue, in cases where the access points are placed in areas where no steady electrical supply is available.

3: Implementations

3.1 Existing Technologies

3.1.1 Wireless Distribution Systems (WDS)

The technology for Wireless Distribution Service already exists, though public adoption is still relatively scarce. WDS creates a wireless backbone between access points in the same wireless network. It functions similarly to a traditionally wired network. New IEEE specifications allow for static routing of data though such a network.

A WDS devices acts in two ways, either as a standard access point for wireless devices, a bridge connecting access points together, or both. DD-WRT, a highly customizable firmware for the WRT-54GL allows WDS, and supports mapping. This allows a cyclic loop of WDS stations without a data cascade, meaning that data will not infinitely travel though the network. Looped networks offer fault tolerance, so that a dropped link will not cause a network failure.

WDS halves the network bandwidth for each access point added to the system, as each signal needs to be received then retransmitted. Clever channel distribution and multiple gateways can reduce the total amount of network traffic.

3.1.2 Antennas

Existing antennas offer great flexibility. Commercially available antennas sport high gain and low loss transmission, highly controlled fields of transmission, and high prices. The same kinds of antennas used for cellular communication are available for Wi-Fi along with many other signal bands, but are out of a low-end price range. [3]

3.1.3 Multiple Gateway

Multiple access points to the greater internet can provide broad ranges of data flow. Should the network become clogged, data can be routed through the best point of access, or if the network is free and open, traffic can be sent to all gateways in order to accelerate downloads and uploads. [2]

3.2 Hardware

The most important part of this project is the hardware, without which there would be no access point. This project is aimed at lowest-cost equipment in order to facilitate installation and implementation.

3.2.1 Access Points

The Linksys WRT 54-GL router is an off-the-shelf unit retailing for extremely low prices compared to high-end commercial electronics. It is sold as a household wireless access point, firewall, gateway, and NAT server all in one box. It is highly customizable, running Linux firmware and sporting detachable antennas. Power consumption is relatively low.

3.2.2 Firmware

At least a dozen versions of firmware exist for the WRT 45G-L. Several versions are listed here. All support WDS, NAT server, simple Access Point, and Gateway mode, and can selectively activate or deactivate each feature. The default firmware does not provide WDS support yet, however it allows all other modes. The default firmware can be used in a pinch if there is no available way to flash new firmware. The DD-WRT firmware allows adjustment of broadcast strength, allowing a reduction in power consumption while still maintaining a clear signal. It allows tree based routing and static routing methods. Can be configured to run without maintenance, in a "fire-and-forget" fashion. It also can be configured to ghost other access points, so that as far as any one wireless device can tell, any access point is the same as another. [5] Tomato offers a better user interface, relatively easier to use. Capable of plug-and-play. [6]

3.3 Unit Design

The design of each unit is the ultimate decided in power consumption. The better the design, the lower the power demand.

3.3.1 Dual Unit

At each physical location, two network devices can be used. Each WRT-54GL has two antennas, and with two devices a total of 4 antennas can be delivered to each location. One device can act as a WDS, using a specific channel, connecting the two neighboring nodes and the other local WRT device. The WDS device can have two highly directed antennas, linking up and down the chain to other WDSs. The other local device, acting as an AP, is fitted with two biquad antennas. This makes it a focused antenna, nearly 180 degrees.

3.3.2 Single Unit Repeater

Alternately, other types of APs can be created for less cost but less backbone support. A single Linsys device at a location can be fitted with two biquad antennas and have it act as a repeater, or to simply "turn the corner" around an obstruction.

3.3.3 Transmission

Extensive research on antenna optimization exists, though are not applicable to this project. The purpose of this project is to keep costs as low as possible, and homemade antennas are not as finely-tuned as professional equipment. The exact measurement required for perfect transmission is not achievable on homemade equipment.

Clients attempting to access the network can connect to these devices, and the overall network coverage can be increased if the clients also enhance their signals with directional antennas. Network bandwidth can be increased if the AP and WDS act on different channels. The AP and WDS devices can be wired, their distance apart is negligible, about 2 inches.

3.4 Channeling

Using the Wi-Fi standard 3 non-overlapping channels, cross-talk between networks can be nearly eliminated. If they are devoted properly using coloring schemes or just foresight, network volume can be tripled. [3]

4 Issues

4.1 Power

Academic papers describing battery power supplies are limited. There are many discussing theoretical power and power supplies, along with ways to generate power. Using batteries, solar panels, and AC power to power the devices is not a matter of academic research, rather a matter of applied electrical engineering. [3]

4.1.1 Batteries

Battery draw and recharge is the critical issue for these units. Each location requires at least 12 hours of battery life, and this may end up exceeding the price of the access points. Fortunately, battery technoogy is improving, and as the country becomes more energy aware, power consumption is dropping.

4.1.2 Solar

Solar power technology has unfortunately stagnated, and is not improving at any significant rate. However, the amount of power required to run the devices is relatively low, less than a half amp, and solar power is plenty capable of generating low-current power. It is a viable option for these access points.

4.1.3 Wind

Small wind turbine have been in use on private sailboats for years. They are compact, lightweight, and efficient, but they require facing the wind and pricing is not very low.

4.1.4 Grid

Taking power directly from electrical transmission lines is the most obvious way. Unfortunately, this project is targeting areas with no or limited power supply. This could

still be tapped in the event that the power is intermittent and can be used to charge the batteries.

4.2 Antennas

4.2.1 Omnidirectional

Omnidirectional antennas are the most basic, simplest, and most common available. They cover a 360 degree wide range and offer effective gain in a close area. Traveling beyond the transmission range causes signals to drop precipitously until they fade to near zero. They are only effective at close range. [8]

4.2.2 BiQuad

The Bi-Quad antenna is a basic ground plate with a quarter, half, or full wavelength wire receiver. They offer moderate to high gain, around 10 dBi if built correctly, around 5 if there is a construction problem. They offer nearly 180 degree spreads and offer a longer range but a geographically separated area. [7]

4.2.3 Yaggi

Directional antennas, particularly D.I.Y. directional antennas, have been known to reach 10 mile links. This is extremely useful to create large-scale networks, however, testing and development is complicated, isolating the units to test are hard to do with signal range exceeding several hundred feet. Also, links shorter than this, which will be the general case, will cause noise at great distances and cross talk. It does also allow for redundancy, a node failure will not be a problem if the devices can simply talk over it. [8]

4.3 Wiring

Some devices local to a transmission point can be hard wired. This can be used to update firmware, or to link two devices in the same location. It can also be used to connect tiers in order to not pollute yet another channel for transmission in order to keep the sections isolated.

4.4 Weather

This is the biggest hurdle. Each component needs to be secured from Mother Nature, including extreme heat and cold, rain and drought, humidity and ice. Lack of wind and sun can prevent recharging of batteries.

5 Future Work

First and foremost, a proof of concept WDS system with two units must be established. Power supply at this point is irrelevant, as the first system should be designed to test bandwidth and data-loss over a relatively large connection. These units would employ grid-based power and use two custom made long range antennas to link them, with a short-range omnidirectional to evaluate their local use. All access points are configured to be Wireless Distribution Service points. WDS points allow for a series of wireless access points to act as a single point from the perspective of the

client. The first installed point is at a gateway to the Internet, if necessary or capable. It is equipped with a directional or focused antenna to direct the signal flow towards the coverage area.

The next access point is placed in line with the sightline of the first access point, configured with a WDS system, and a directional antenna back towards the first point, another directional antenna toward the next point. This can be repeated indefinitely into the distance, provided there is enough bandwidth and power. This constitutes a single tier. Provided this works a third unit would be added to the mix in order to measure the long-range effectiveness. This would be an attempt to discover the unit's optimum max-range while maintaining a low signal-to-noise ratio. Once these links are functioning and data-loss and packet-loss have been minimized, power consumption will be evaluated. Batteries, initially simple off-the-shelf batteries, will be used to evaluate their lifespan in operation, then solar cells will be equipped to ensure that a basic recharger can supply the power demands for each unit.

6 Sources

[1] M. Burkhart, "Analysis of Interference in Ad-Hoc Networks", Diploma Thesis, 2003

Jinyang Li, Charles Blake, Douglas S.J. DeCouto, Hu Umm Lee, Robert Morris, "Capacity of Ad Hoc wireless networks," International Conference on Mobile Computing and Networking, Proceedings of the 7th annual international conference on Mobile computing and networking, Pages: 61 - 69, 2001

[2] TeePipe: RAID for Internet Connections

http://josh.com/teepipe/index.htm

[3] LinksysInfo.org, WRT54G.net, (and strangely enough, en.wikipedia.org/wiki/WRT54G)

[4] Aaron Weiss, "Turning a $60 router into a $600 router," November 8, 2005, http://www.wi-fiplanet.com/tutorials/article.php/3562391

[5] DD-WRT: dd-wrt.com

[6] Tomato: http://www.polarcloud.com/tomato

[7] Trevor Marshall, "BiQuad 802.11b 11dBi wide band antenna," http://www.trevormarshall.com/biquad.htm

Martin Pot , "Wireless Networking Info, Biquad Antenna Construction," 22 November 2006, http://martybugs.net/wireless/biquad/

[8] Qunfeng Dong, Yigal Bejerano, Suman Banerjee, "Building Robust Wireless Mesh Networks Using Directional Antennas: How Many Radios Are Enough," http://www.cs.wisc.edu/~qunfeng/papers/UW-CS-TR1572.pdf

[9] Su Ui, Yong Pei, Shivkumar Kalyanaraman, "On the capacity improvement of ad hoc wireless networks using directional antennas," International Symposium on Mobile Ad Hoc Networking & Computing, Proceedings of the 4th ACM international symposium on Mobile ad hoc networking & computing, 108-116, 2003

Wireless Security: Secure and Public Networks

Kory Kirk

Villanova University
Computer Science
kory.kirk@villanova.edu
www.korykirk.com/

Abstract

Due to the increasing amount of wireless access points that provide service for the general public in places like hotels, restaurants, airports, train stations and cafes, there is a need for a wireless protocol that allows public access to secure wireless access points. Security protocols exist for wireless networks; however, all widely implemented protocols require some sort of authentication that is unavailable to public users. Few attempts at implementing a protocol which allows public access to a secure wireless network have been made. In this report we discuss these protocols and propose new solutions.

1. Introduction

Wireless networks are everywhere, and ensuring secure wireless communication is of great importance. This research investigates infrastructure networks. This kind of network is based upon one or more connected Access points (AP) that end users connect to in order to access the network. Security for infrastructure wireless networks is usually ensured through the use of a wireless portal or the implementation security protocol. A wireless portal is a computer running server software that manages every connection on a given access point. A security protocol is a standardized software package that is installed both on an access point and the client computer. This software dictates how the computer and access point should communicate.

The problem addressed in this research is the lack of a common protocol that provides a secure connection to an access point, but does not restrict access to that network from the public. A wireless network using WEP is secure, but restricted, meaning that people need to know a secret key in order to connect to the network. The main use would be in public places. Consider a wireless network in an airport - in order for it to be public, it has to be non-secure, which puts users at risk. We investigate alternatives which would enable the implementation of a public and secure wireless network. Does a protocol like this already exist? Are there implementations of current protocols that could ensure a public and secure network? If these do exist, why are they not widely used?

2. Some Known wireless vulnerabilities[1]

In order to better understand wireless security and what is needed in order to create a secure and public network, it is important to know vulnerabilities that are common to wireless networks.

2.1 Wireless Sniffing/Wireless Probing

Wireless sniffing is a method of wireless hacking in which a computer picks up packets of information that are being sent between wireless clients and access points. These data packets could belong to a secure or insecure communication session. “Even though the attacker gathers considerable amount of information regarding a wireless network through sniffing, without revealing his wireless presence at all, there are pieces that may still be missing. The attacker then sends artificially constructed packets to a target that triggers useful responses. This activity is known as probing or active scanning. “[1]

2.2. MAC Address Spoofing / IP Spoofing

The *MAC address* is a unique identifier burned into the network card, wired or wireless.

An *IP address* is a unique identifier associated with a computer within a network. Unlike a MAC address, the IP address can change from network to network.

When intruders are able to spoof their IP address or MAC address, they are changing the network fingerprint of a computer, therefore allowing their computer to be viewed as any specific computer on the network. This turns out to be very dangerous, especially in wireless networks, because with many implementations of wireless portals, access is provided only to computers whose MAC addresses are recognized. This allows other people to sniff out packets between an access point and a client, and decode their MAC address, spoof it, and basically become that computer on the network.

IP Spoofing happens when a computer sends packets to another computer and spoofs the source of the packet. In such situations the receiver will believe that it is coming from the access point, but it could have easily been sent by another computer that spoofed the IP address.

2.3 Man in the Middle Attack

The *Man in the Middle* (MITM) attack is a very serious attack, in which an “attacker can merely observe the communication or modify it before sending it out. An MITM attack can break connections that are otherwise secure. At the TCP level, SSH and VPN, e.g., are prone to this attack[1].” In an MITM attack all data that goes from the client to the access point and vice versa also goes through the attacker.

The attacker that is performing an MITM uses two network cards. One of these cards is to act as the access point to the client, and the other card is to act as the client to the access point. The attacker first sends a Deauthentication frame to the client using the access point’s MAC address, so that the client believes that it has been disconnected from the access point. At this point, the user then begins to search for the access point in order to reconnect to it. The attacker’s network card that is acting like an access point is broadcasting the same SSID (Service Set Identifier, the actual name that an access point broadcasts so that users know what network

they are connecting to) and has the same MAC address as the access point. At this point in the attack, the client sees the legitimate access point and the attacker's network card as identical access points. The user's computer chooses which access point to connect to based upon which one is broadcasting a stronger signal. If the user connects to the attacker's fake access point, the attacker then forwards all the information from the access point to the user and vice versa. This allows the attacker to modify any information before forwarding the packets. This kind of attack can also be initialized by a technique called Address Resolution Protocol (ARP) poisoning.

ARP is a protocol that enables a network to find a user's hardware address (MAC address) when the network has only the user's IP address available. When a network receives a packet whose destination IP address is unknown to the network, ARP caches the packet in a queue and sends an ARP Request packet over the subnet asking if any machine's IP address matches the unknown IP address. If there is a matching IP address, that machine responds with an ARP Reply packet, containing its MAC address. The ARP client table (the table that holds all of the MAC and IP addresses) is then updated and the packet is dequeued and sent to the destination.

ARP updates its table if it receives an ARP Reply packet, regardless of whether it is in response to an ARP Request packet or not. This means that an IP Address can be matched with a new MAC address whether or not the source IP address stored in the ARP Reply packet is in the ARP table. This allows *ARP poisoning*, which happens when an attacker spoofs its IP address and sends an ARP reply packet to a machine with the wrong MAC address in order associate that MAC address with the spoofed IP.

In an MITM attack using ARP poisoning, the attacker performs ARP poisoning attack on both the access point and the user. This makes the client believe that one of the attacker's network cards is the access point, and it makes the access point believe that the other network card is the client. From this point on, all the attacker needs to do is forward the information from the access point to the client and vice versa, modifying any information that he/she wants. ARP poisoning is a more effective way of initializing an MITM attack.

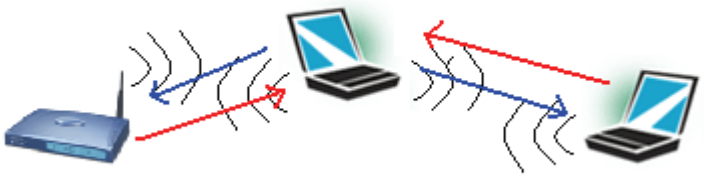


Figure 1 (MITM Attack)

2.4 Denial of Service

A *denial of service* attack on an access point happens when an attacker uses multiple network cards to send a large amount of packets to the access point in an attempt to block any other information from getting in. These unauthorized clients flood the outlet a legitimate user

may want to connect to, but cannot connect due to the high amount of resources that the unauthorized users are consuming.

3. CHOICE Network[2]

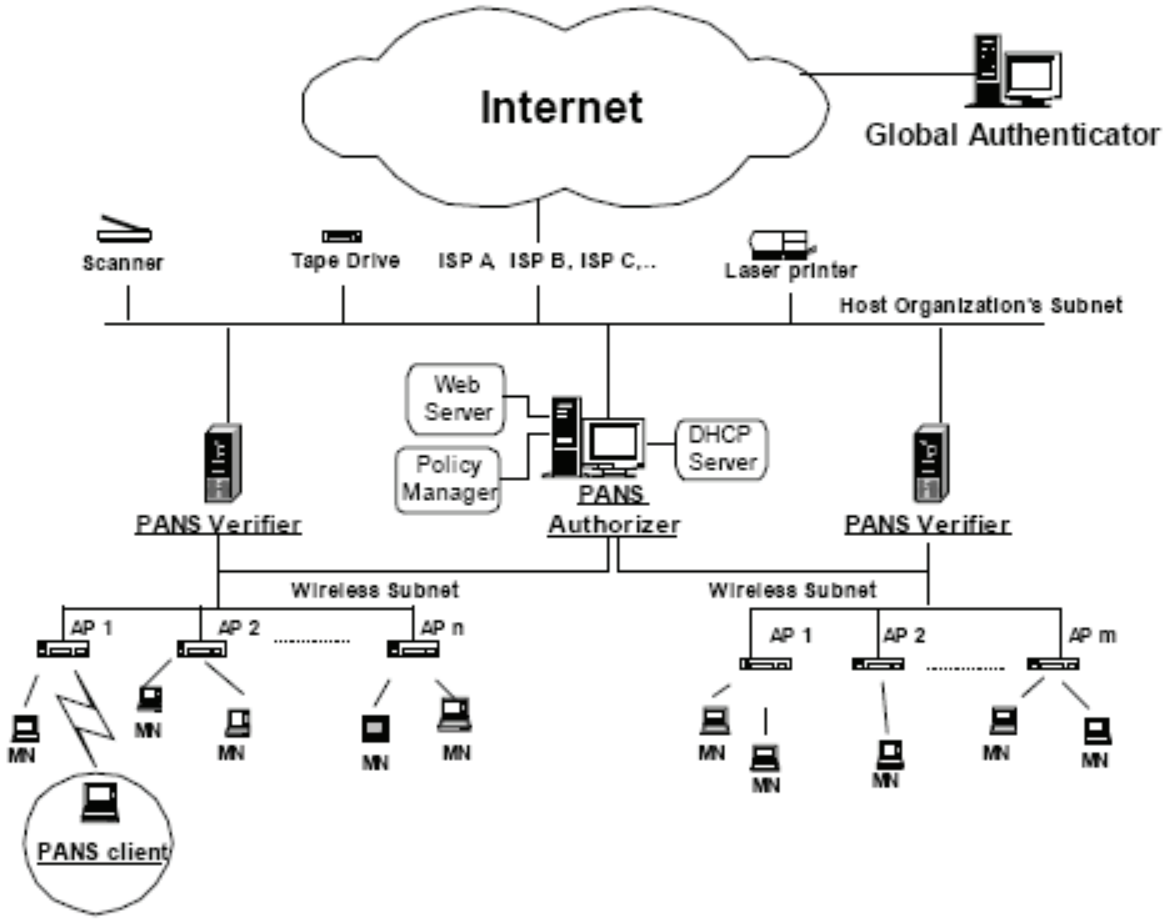


Figure 2 (Choice Network[2])

3.1 Implementation of the CHOICE network

3.1.1 Protocol for Authentication of Negotiation of Services (PANS)

"PANS facilitates authentication, authorizes access, enforces policy and last hop QoS[Quality of Service], and provides privacy to network users and accounting to network operators[2]." PANS is the key security protocol which is the basis for Microsoft's CHOICE network. It comprises the software that is installed on the end user's computer as well as on the PANS verifier, which is an important part in authentication in the CHOICE network.

3.1.2 How it works

Refer to Fig. 2 throughout this section. The client first connects to the public access point, and receives an IP address. Using HTTPS (secure Hyper Text Transfer Protocol), the user is able to download needed software from the internet in order to complete the PANS authentication. From here, the user authenticates with the PANS verifier, which verifies the authentication request, and then sends it to the Global Authenticator. The Global Authenticator then returns to the PANS client and verifier a key and a token. This key and token are used to encrypt and authenticate all data that will be subsequently transferred between the client and the access point. In addition, the PANS verifier assigns a key identifier which is the index of the (key, token) pair stored in a table by PANS verifier.

The PANS verifier accepts DHCP (Dynamic Host Configuration Protocol that assigns IP addresses to clients), PANS (secure data between the access point and users), and HTTP packets which are part of the initial authentication process. This restriction of packets secures the subnet so that only secure data or data that is part of the authentication process may get through.

After authentication, all data sent between the access point and the client is encrypted using the key and token assigned to that specific user. It is then decrypted by the verifier and forwarded to the destination. The key and token are only sent from the access point to the client once using HTTPS, so it is encrypted data. From that point on, only the key id is sent with each packet, and the PANS verifier uses that id to look up the key and token in order to decrypt the data. Because it uses this key id to represent the user instead of an IP address or MAC address, it makes it less vulnerable to IP or MAC address spoofing, which includes MITM attacks.

3.2 Possible Applications to a Public and Secure network

The CHOICE network provides a few things that would be useful in implementing a Public and Secure network. Actually, it would be an ideal implementation for a public and secure network if it were not for the Global Authenticator. In this case, the user is required to have a login and password in order to have access to the network, which makes it restricted. If the key and token were assigned by the PANS verifier without having to do some sort of user authentication, then users could have their own key and token to uniquely encrypt their own connection.

3.3 Potential vulnerabilities

Even though the key and token are initially sent using HTTPS, there is still a possibility that an MITM attack could be taking place during this transfer, and the attacker could then decrypt the encrypted HTTP packet and retrieve the key and token. Although it would be difficult for the attacker to decrypt the data using the key and token from there, it is a possible circumstance nevertheless. Either way, it provides some level of security compared to a public network.

4. *Linux Wireless Portal[3]*

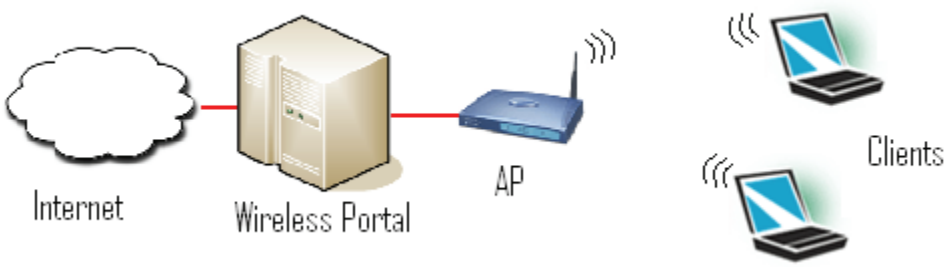


Figure 3 (Wireless Portal)

4.1 Overview

A Linux Wireless Portal (see Figure 3) is a machine that is connected to an access point. The access point forwards all incoming packets to the machine. The machine is running server software that handles all the packets coming from the subnet, and also handles all incoming packets. This allows the Wireless Portal to restrict or allow any access to the subnet that it is managing.

4.1 Implementation

All that is needed in order to implement a Wireless Portal is a machine running the appropriate software. It is pretty simple to create a wireless portal, because it really just acts like a network proxy, forwarding all the information from the client to the destination. And because all data in a network passes through the portal, the portal can restrict access to a certain protocol, or a certain website. The portal can also restrict access to certain IP's or MAC addresses. Common applications used in Wireless Portals are Monowall, noCatAuth, and WiFiDog[3].

4.3 Analysis and Possible Applications to a Public and Secure network

The main problem with Wireless Portals is security. The network packets are just forwarded and not encrypted, leaving them vulnerable to wireless sniffing and probing. The network is also susceptible to MAC and IP address spoofing, especially in a Wireless Portal that restricts access based on MAC or IP address. However, a wireless portal is relatively easy to implement, and does not need a machine with a lot of storage space or memory. Furthermore, wireless portals are not computationally intensive, meaning that they could be implemented with low-end computers.

A wireless portal is a potential choice for implementing a public secure network, since it can basically restrict or allow access in various ways. This allows the implementation of any security measures that are needed. A Wireless Portal is very open-ended and customizable, which makes it a more viable option than a new wireless protocol when thinking about possible implementations of a Public and Secure network.

5. *Proposed Future Work*

We propose to design and analyze the performance of a secure and public wireless network that combines various features of the CHOICE network and Wireless Linux Portals. Our design will consist of using a Wireless Portal to implement something like PANS. Where a computer connects to a

network, receives an IP address, and until that computer is authenticated, no packets will get through to the internet or any other part of the network, except for the portal. The portal would respond to any HTTP requests with a forward reference to a website/web application located on the portal. This web application would force the user to download and install a security application in order for him/her to connect to the internet through the network. Once the user had installed the application, the client computer will send a request for authentication to the wireless portal. The portal will respond with a key and a token (much like the CHOICE network), and a key id which is the index of the key and token in a table stored on the portal. This message will be sent through HTTPS to ensure that it cannot be easily intercepted. From this point on, the user is authenticated with the wireless client, allowing all the data between the portal and the client to be encrypted, then decrypted and forwarded to its destination.

This is the basic idea behind the CHOICE network, but without the Global Authentication. It will be our goal to come up with a detailed design and a plan for implementing such a network, and also analyzing the speed of the network and the resources that it would use up.

References

[1] Prabhaker Mateti , **Hacking Techniques in Wireless Networks**, http://www.cs.wright.edu/~pmateti/InternetSecurity/Lectures/ WirelessHacks/Mateti-WirelessHacks.htm [Date accessed: 12/3/07]

[2] Paramvir Bahl, Srinivasan Venkatachary, **Secure Wireless Internet Access in Public Places** http://www.cs.ucsd.edu/~abalacha/research/paccess pointers/ICC01.pdf [Date accessed: 12/3/07]

[3] Paramvir Bahl, Srinivasan Venkatachary, **Build a Secure Wireless Portal with Linux** http://www.enterprisenetworkingplanet.com/netos/article.php/3592556 [Date accessed: 12/3/07]

Methods for Improving Readability of Speech Recognition Transcripts

John McCoey

Computing Research
Department of Computing Sciences
Villanova University, Villanova, Pa, 19085
john.mccoey@villanova.edu

November 14, 2007

Abstract

With the increasing use of computers as tools for disabled persons to interact and communicate, the need for Speech Recognition software to produce valuable real-time text transcripts has grown dramatically. In order for one of these transcripts to be truly useful, they must be both accurate and readable. The useful transcript should contain both words that are correctly recognized and ideas and thoughts that are maintained. Significant advances since the introduction of commercial Speech Recognition has led to several systems attempting to improve the readability of the resulting transcript. In this paper we review these systems and identify areas where much work remains to be done in order to refine a perfect and accessible transcript for use as a substitute of physically hearing and interacting during a speaker's lecture. We also propose ideas for continuing research, including ways to improve an existing system to become a practical, everyday tool for both the hearing and hearing-impaired.

1. Introduction

The readability of a text document refers to the value of the document for a reader in terms of understanding and communicating the speech. This paper reviews several methods which increase the readability of recorded speech translated automatically into a text document. Section 2 describes the history and background of Speech Recognition, and the main points of focus in our research. Section 3 explains methods for measuring the readability of the transcripts, while Section 4 details various systems which are in use today to automatically increase readability and accuracy. Next, Section 5 lists several ideas for future research directions and improvements for use, especially in a classroom. Finally, Section 6 focuses on our own proposals for creating a formatted version of an existing Speech-to-Text Translation System, while Section 7 is a conclusion of our research.

2. Speech Recognition

Research on Speech Recognition was initiated in 1936 by AT&T's Bells Labs, in an effort to create secure conversation during World War II. Speech recognition did not become commercial until 1982, with the introduction of Covox. By the year 1986, [1] Carnegie Mellon University developed Sphinx, the first system that demonstrated the possibility of a continuous-speech, large-vocabulary recognition system independent of speaker changes. Since then, speech recognition has improved dramatically, and is used in many applications, but it is still not 100% accurate.

Since the introduction of Covox in 1982, Speech-to-Text Translation Systems have been used in a variety of situations, including automated telephone voice menu systems, real-time dictation and transcription for word processing, and assistive technology for people with disabilities [2]. However creating a perfect Speech-to-Text Translation is a difficult problem to solve, due to a vast pool of words and phrases that a computer must understand, in addition to a wide range of languages, accents, and other voice characteristics.

This report summarizes research on techniques that attempt to make text translations of spoken words more precise, while reducing cost and time, and retaining the readability. An ideal translation of Speech-to-Text could potentially solve the problem of phone communication for individuals who are hearing-impaired as well as allow speeches, classroom lectures, interviews, or court cases to be documented without a specially trained stenographer for use by both the hearing and hearing-impaired alike. However, in order for these scenarios to be effective, the final transcripts require a high level of accuracy and readability.

3. Measuring Readability

Readability of a transcript is important for the reader in order to understand and comprehend the full text as well as if it were spoken into the computer. However, when a person speaks, things such as punctuation and changes in train of thought are not spoken, and must be filled in by the translator. Similarly, pauses and phrases that are not necessary can be omitted to increase readability, but a system must careful not to remove necessary words. Often, the best way to ensure high readability is for the speaker to speak carefully and grammatically correct. This option, however, is often not practical in real-life situations. There are many Speech-to-Text algorithms that attempt to increase readability, which will be discussed throughout this research paper.

Ross Stuckless [3] argues that word "accuracy" is important, but is not the best way to measure one's ability to comprehend the transcript. This is why Stuckless developed the readability percentage test in order to compare the readability of transcripts from different systems. Stuckless' algorithm requires someone grading the Speech-to-Text Transcript to follow the steps below while simultaneously reading the transcript and listening to an audio version of the speech:

1. *Add the number of speaker changes, total words spoken, and total sentences spoken.*
2. *Add the number of times the identification of speaker changes were omitted from the text, total word errors, and the number of times sentence identification was omitted.*
3. *Subtract (2) from (1).*
4. *Divide (3) by (1) to obtain the readability percentage or score* [3].

As Jones, et al., [4] explains, word accuracy is very important to the overall readability of a document, as is evident in their experiment which shows that reference texts, both edited and unedited, have a higher readability than system produced transcripts. A reference text is a transcript written by Jones et al. using a theoretical upper bound for word accuracy of a Speech-to-Text system. Even the standard reference text provides the reader a higher understanding than the cleaned up version of the system texts, which when done by the system, actually seems to decrease readability.

Jones, et al., however, use a different method [5] of measuring readability using a case study. This method asks participants to read the translated texts and answer questions, and then determines the readability based on the amount of time taken to read the text and answer the questions as well as the accuracy of the answers given. Just as there are several ways to measure the readability of a document, there are several different systems and implementations of increasing readability, none of which are perfect.

4. Speech-to-Text Translation Systems

The various models of Speech-to-Text Translation Systems in use today, all of which use different methods to increase the speed and word accuracy, create an efficient and accurate transcript which can be easily post modified for a highly readable result. Some examples of such systems include Sphinx, a Lexical Search Tree, and the VUST model. Sphinx [6], the most widely used of these systems, uses an efficient and fairly accurate method. Another theoretical example [7] by Mosur Ravishankar is difficult to implement, but uses a lexical tree along with post-processing to create a system which could potentially create higher accuracy results at a much faster rate compared to the Sphinx system. The VUST [8] system has been developed at Villanova University and builds upon the Microsoft Speech Recognition Engine, or MSRE. The MSRE is built into all computers running the Windows operating system.

4.1 Sphinx

Sphinx is one of the first speech recognition systems created. The Sphinx system was invented in 1988 at Carnegie Mellon University and became the first high performance Speech-to-Text Translation System [9]. The system was the first to use the idea of Hidden Markov Models (HMM) to choose the word with the highest probability match for speech sounds. While the original Sphinx is no longer used, as developments have led to several updates, the latest of which is a modern version called the Sphinx IV

system [6], the theory behind the first Sphinx system remains vital to understanding how Speech-to-Text Systems work.

The main concept behind Sphinx is the Hidden Markov Model, or HMM: a mathematical model for generating probabilities of change from one state to another. We have created a very simple example of an HMM which uses the probabilities of weather as shown in Figure 1.

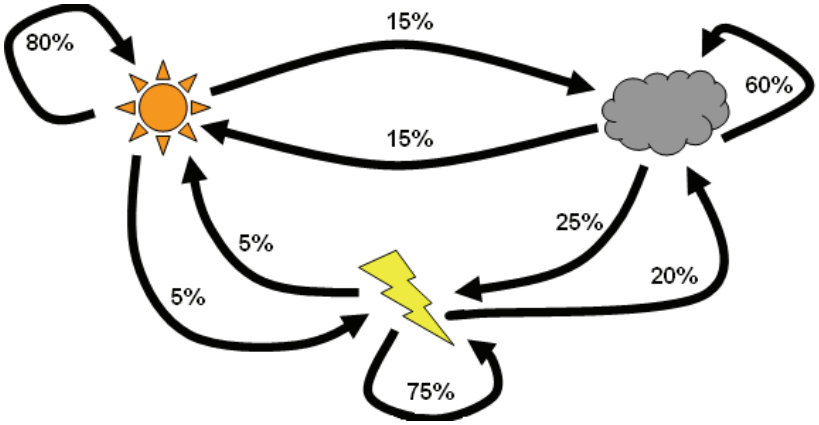


Figure 1. HMM of Weather Probabilities

We can see in Figure 1 that if the current weather is in a sunny state, there is an 80% chance that it will remain sunny, and a 5% chance of a thunderstorm. However, if it is cloudy, there is a 25% chance of a thunderstorm and only a 15% chance of sunshine, and so on. There are also hidden states which are not shown, such as the probability of snow or clear skies. This type of model is used in speech recognition software by creating a model of probabilities which simulate generalized triphones. A triphone is "simply a distinct phonetic unit labeled with its immediately adjacent phonetic contexts. Triphones were, and remain, one of the most effective innovations in modeling speech sounds in HMM-based systems" [9].

An HMM attempts to determine, based on the syllables already spoken, what the next syllable will be. This is then applied to words as well, as the first Sphinx system kept a library of word pairs which were allowed by the language, and which were not. The HMM would determine the highest probability for the next word based on the probability of which words had followed the word spoken in the past. This makes it easier for the HMM to learn how a speaker creates sentences and phrases, as well as which words are most likely to be used in a given situation [6].

The HMM system used by Sphinx ensured a very high accuracy rate for the first time in the history of Speech-to-Text Systems with a large vocabulary. According to Singh [9], for most lectures using a 1000-word vocabulary, the Sphinx system demonstrated a 90% accuracy rating. It was also very efficient, because Sphinx could be run in real time, displaying the words as they were spoken.

4.2 Lexical Tree Search

A theory proposed by Mosur K. Ravishankar of Carnegie Mellon University suggests that a Lexical Tree Search [7] may surpass the HMM as the future of Speech-to-Text Translations. A Lexical Tree provides a way to reduce the search space for a Speech-to-Text System word search by combining similar sounding pronunciation keys into a tree. This avoids having to save duplicates for words or branches of words that sound similar, when they can all be stored into an easily searchable tree. This method is thought to be as much as 5 times faster than the Sphinx system, but because of the combination of words, it loses some accuracy without the probability of the HMM. When used with post processing, this approach can yield the same fast speed while increasing the word accuracy to a level very close to that of the Sphinx system [7].

Mosur Ravishankar [7] suggests post-processing solutions that would improve the accuracy of the Lexical Tree Search algorithm, which include employing algorithms such as the Phonetic Fast Match Heuristic, Re-score Lexical Tree Word Search, and the Global Best Path Search in addition to the Lexical Tree Search. The Phonetic Fast Match Heuristic looks ahead at the next several basephone roots of the lexical tree and eliminates those phones which are not currently active. This in turn eliminates words that can not possibly be spoken based on the first syllable, and doubles the speed of the final search. After the search, the lexical tree outputs a lattice of the words found. This can be re-scored to create a list of the most likely word choices based on pronunciation and context. Finally, the Global Best Path Search is the hardest to implement. Sphinx II uses Viterbi Search [6], which determines the probability between two word choices based on the single word spoken immediately before and immediately after the word in question.

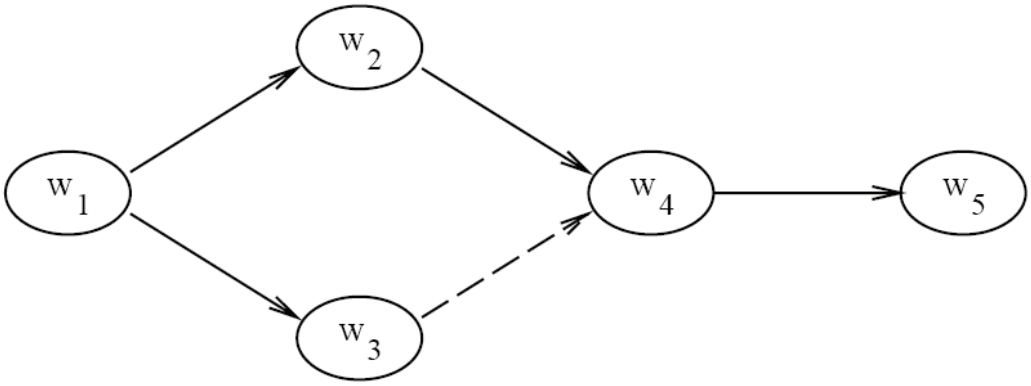


Figure 2. Sphinx-II Viterbi Search [7]

This approach is illustrated in Figure 2. Global Best Path Search [7], on the other hand, looks at all spoken words to find the most probable word selection. For example, in Figure 2, each node represents a possible word choice for the Sphinx-II system. Given the words chosen for w_1 and w_4, the probability values suggest that w_2 is a more likely choice than w_3, so w_2 is chosen and the system moves on. In Global Best Path search,

however, if the system reached w_5 (or even w_6, w_7, $w_{8...}$) it may realize that w_3 was in fact a more likely choice, and will backtrack and fix this error. Because it is not limited to the words immediately before and after the word in question, this process requires a second pass through the text after all the words have been spoken in order to obtain a true probability of each word choice, which we find impossible to implement in a real-time situation [7].

4.3 VUST

Kheir and Way [8] created the Villanova University Speech Transcriber, or VUST, an automatic speech transcription system which enhances the Microsoft Speech Recognition Engine (MSRE) to create fast and accurate transcriptions. The VUST System works with two additional pieces of software, a Training Engine, and DiBS. Information from this software is used to increase the vocabulary and the accuracy of the MSRE.

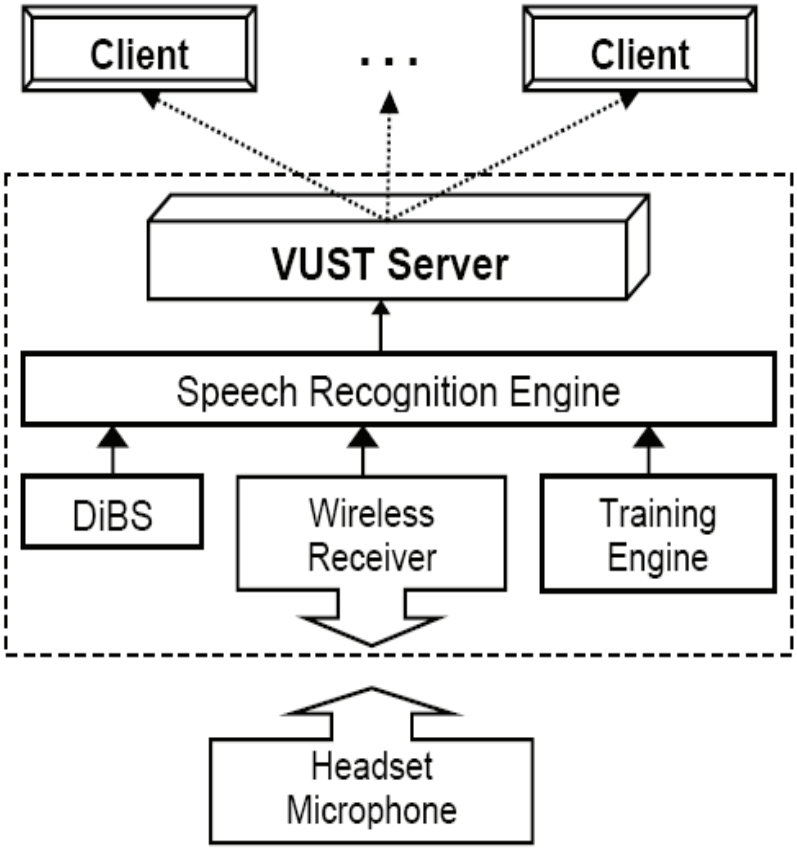


Figure 3. VUST System Design [8]

The VUST system is depicted above in Figure 3. The dotted line around the center portion of the image represents the computer that the VUST system is running on. The headset microphone sends information to the wireless receiver, which forwards it to the Speech Recognition Engine along with the information from DiBS and the Training Engine. Theoretically the DiBS and Training Engine software components could be used

in conjunction with any Speech Recognition Engine, such as Sphinx or the Lexical Tree Search which we discussed earlier, but in this particular case, it is the MSRE. The transcript from the Speech Recognition Engine is then put on the VUST server, where it can be downloaded to clients, as seen in Figure 3.

Ultimately, the goal of the VUST system would be for the server to exist outside of a personal computer, and allow students to simply download the transcript from anywhere on campus directly from the server. Ideally, the server would also be able to store information about particular speakers, such as the results of the training engine, and would be completely invisible to both the speaker and reader [8].

4.3.1 DiBS

One of the most important applications created for use in a university classroom is VUST's DiBS. DiBS stands for Dictionary Building Software, and is a key component of the VUST system which allows for an expandable and topic specific vocabulary to be implemented for the Speech-to-Text transcription. DiBS allows the user to add custom words to the recognizable vocabulary of the Speech-to-Text System. These words can be added in a variety of ways, including scanning a given document for larger words that it does not recognize, and adding them to the dictionary. This is essential for applications such as teaching Computer Science classes where the language is constantly changing [8]. The words added to DiBS are added to the MSRE using rules for pronunciation of letters in order to make the best assumption of when a word should be recognized. If the words added through DiBS are still not recognized, either because of exceptions to the typical pronunciation rules or complicated spelling, they can be specifically trained using the VUST Training Engine.

4.3.2 Training Engine

The VUST Training Engine requires the speaker to read a list of words and phrases into a microphone for approximately 30 minutes in order for the system to adapt to recognize patterns of speech by the particular speaker. This is an optional step in the VUST system, but increases the accuracy of the transcriptions, and needs to be performed only once for a speaker before infinite translations are made [8]. The Training Engine also gives an option to read words which were added into the DiBS aloud, and train the system to better recognize the new vocabulary as well. According to Kheir and Way, when DiBS, the Training Engine, and training of words added through DiBS are used, word accuracy of computer science lectures increased to near 94%, a huge increase from the 75% accuracy of the untrained MSRE [8]. This setup was tested in a classroom setting at Villanova University, which was found to be very helpful to both hearing and hearing-impaired students, yet many steps still remain to create a completely accessible and practical option for widespread use in university classrooms.

5. Ideas for Future Research Directions

There is still much to be explored in the task of improving the readability of Speech-to-Text Transcripts, especially for use in the classroom. Some of the main ideas revolve around increasing the word accuracy of the system itself. This is a huge step in achieving a readable and helpful transcript because regardless of structure, if the words are wrong, reading becomes difficult. The next step is to show indications of grammar and structure throughout the transcript to make reading the transcribed words more natural. This is a difficult task because humans rarely speak in the same way in which they write or read. Finally, incorporating these systems into classrooms is a difficult task, but one that would make learning and participating in a classroom accessible for hearing-impaired students [10].

5.1 Classroom Use

The VUST System has already been tested in a classroom [8] as a projection of the words spoken by the professor in real time during a computer science lecture. All of the students, including those who were not hearing-impaired, found the output to be of enormous benefit, and agreed it matched generally what was taught in class. However, there is still a need to be in class, as the transcript does not offer the full benefits of being physically present. Wald, M., Bain, K. and Basson, S. H. [11] detail a new movement in an effort for the hearing-impaired to have Automatic Speech Recognition systems in classrooms through the Liberated Learning Project (LLP). LLP [12] looks to use new technology to even the playing field for disabled and no disabled students by making classes accessible to all students with devices such as Speech-to-Text Systems. While these are all great improvements to the classroom for hearing-impaired students that exist currently, there are still major issues that need to be fixed to increase the accuracy and readability as well as the accessibility of these systems [10].

5.2 Increase Accuracy and Readability

Continuing research to increase the accuracy and readability of Speech-to-Text Transcripts must increase the amount of information that can me gather by simply reading the transcript. One idea presented by Wald, M. and Bain, K would be to include indications of emotion, mood, and tone of voice [10]. This would greatly increase the ability to detect anger, sadness, sarcasm, and questioning phrases which may have been too subtle to pick up on otherwise.

Jones, et al. [5] look to research which errors are the most harmful for the readability of a transcript, and to make correcting those errors a priority over less important errors. For example, capitalization of the first letter of a sentence may not be as important as indicating a change in speaker. To increase the accuracy of Speech-to-Text Transcripts, Wald, M. and Bain, K. also suggest two ideas based on word confidence and phonetic spelling [10]. For instance, if a word is not recognized with a high enough probability percentage for the computer to be sure it has chosen the right word, it could give the top 3 choices it has with the probability rates it calculated. Or, the document could also print

out the phonetic spelling of syllables that make up the mistranslated word and allow the reader to make their own translation of the word in question.

5.3 Increase Accessibility

The accessibility of Speech-to-Text Systems, especially in classrooms, is a difficult problem to solve. While projecting a transcript on a screen in real-time during a classroom may seem like an easy solution, it has problems too, because it requires a hearing-impaired student to constantly read the screen, which takes away from other things that may be going on in class at the same time, such as notes being written on the board, or a professor solving a problem or explaining diagrams which require the visual attention of the student as well.

Kheir and Way [8] suggest an expansion of their DiBS tool could also help accessibility because dictionary classes that apply to certain classes or certain teachers could be saved to a server and uploaded at the appropriate time. Also, a personal profile could be saved for each teacher which would remember the results of the Training Engine and would automatically load every time that person speaks to the software. This eliminates the need to re-train the software every time a speaker or class is changed [8]. Another way to minimize the problem of a student needing to look at two screens during class, or have two separate documents when studying for a test, is to intermix the translated speech directly into a document which is being discussed and explained to show where a person is talking about [10]. Another major issue in accessibility, even for those who are not hearing-impaired, is the difficulty to translate spoken word into different languages, for use in other countries, or in multilingual classrooms [5].

6. Proposed Work

Past research has shown that the key to improving accuracy, readability, and accessibility of Speech-to-Text Transcripts is to change the single sentence output of a translation system into an organized paragraph of sentences with proper punctuation.

We propose to design and analyze methods which would accurately formulate visual indications of pauses, sentence structure and punctuation, and changes of speaker in a paragraph form. This is the step that would make a transcript seem more like a formal written paper rather than a jumbling of words in a large unorganized paragraph. In order to do this, we need a system which does not already make adjustments based on speaker pauses. In our case, the VUST System [8] is the most accessible, as it is both free and easily upgradeable. We propose to modify the DiBS in order to recognize speaker pauses and the approximate length of time of the pauses, and add them to the translation. For example, depending on the length of time a speaker stops between words, we intend to create specific symbols representing short pause, pause, or long pause which will appear in the text where the pauses occurred.

We also propose to implement another software device, in addition to the Training Engine and DiBS, which would format the resulting transcript into paragraph form, based on the pauses dictated. This would require research into the average length of a speaker's pauses in various situations. We propose to research various speakers and implement a system which determines a relative ratio of the difference in pause lengths, rather than a simple scale based on seconds. However, every speaker will pause for different lengths of time, depending on how fast or slow they speak, so some of our information about relative pause length must come from the supplied Training Engine.

The software device we plan to implement would in theory read the entire transcript including the pause length symbols and format a paragraph from the results. For example, short pauses may indicate the need for a comma, while a normal pause could be the end of a sentence. In this case a period would replace the symbol for pause and the next word would be capitalized. A longer pause might indicate a new paragraph or a loss in train of thought. This could also be used to determine a change in speaker, as if after a long pauses, the length and frequency of pauses changes, it may indicate a new person has begun speaking, which could be noted as well.

7. Conclusion

Since the first commercial use of Speech Recognition in 1982, the systems of Speech-to-Text Translation have made huge leaps in accuracy; however, the readability and accessibility of these systems still suffers. We believe our research in this area should look to improve the ways in which word accuracy is achieved, as the value of achieving the correct words cannot be made up with the addition of punctuation and indications of speaker pauses. Also, we emphasize that the most important step in all of this research is the increase of accessibility. There is no use increasing the readability and accuracy of Speech-to-Text Transcripts if they will not be used. These systems must be easy to implement, and demonstrated throughout our everyday lives. In order for this to happen, we must research especially ways to detect different speakers and multiple languages because that is what will allow for useful access for the entire world.

If systems of Speech-to-Text Recognition remain as they are today; highly accurate only when detecting a single language from one single speaker, who specifically trained the system, it will fail as an unpractical solution. Luckily, we believe advances to both software and hardware will greatly help the case for these systems in the near future. With continuing research in manufacturing hardware, faster and cheaper computers should allow for much more processing to be done in a short amount of time in the future. This could yield much higher readability statistics as systems could process grammar and word pauses into the transcripts and without wasting much time.

We can also see that the accessibility of these systems looks to improve as well, with projects such as LLP [12] striving for fair opportunities for disabled persons. In the United Kingdom already, there are laws requiring the funding for use of Automatic Speech Recognition systems for any class with a hearing-impaired student [10]. This could soon be true for the United States and other parts of the world as well as the

software becomes more practical and less expensive, and the demand for fair opportunity of hearing-impaired students grows.

References

[1] Markoff, John; "Talking to Machines: Progress Is Speeded"; The New York Times, Business Technology; July 6, 1988.

[2] Terry Thompson; "Tech Tips: Are You Talking To Your Computer Again?"; Disabilities, Opportunities, Internetworking, and Technology, University of Washington; 2006.

[3] R. Stuckless; "Recognition means more than just getting the words right: Beyond accuracy to readability"; Speech Technology, Oct. /Nov. 1999, pp. 30—35; 1999.

[4] Jones, Douglas, et al.; "Measuring the Readability of Automatic Speech-to-Text Transcripts"; Proc. Eurospeech, pp. 1585-1588; 2003.

[5] Jones, Douglas, et al.; "Measuring Human Readability of Machine Generated Text: Three Case Studies in Speech Recognition and Machine Translation"; MIT Lincoln Laboratory, Department of Brain and Cognitive Sciences; 2005.

[6] Willie Walker, Paul Lamere, Philip Kwok, Bhiksha Raj, Rita Singh, Evandro Gouvea, Peter Wolf, Joe Woelfel; "Sphinx-4: A Flexible Open Source Framework for Speech Recognition"; Sun Microsystems Inc., Tech Rep.; TR-2004-139; 2004.

[7] Mosur K. Ravishankar; "Efficient Algorithms for Speech Recognition"; Ph.D. Thesis, Technical Report CMU-CS-96-143; Computer Science Department, Carnegie Mellon University; 1996.

[8] Richard Kheir and Thomas Way; "Inclusion of Deaf Students in Computer Science Classes using Real-Time Speech Transcription."; ITiCSE'07; Applied Computing Technology Laboratory, Department of Computing Sciences, Villanova University; 2007.

[9] Singh, Rita; "The Sphinx Speech Recognition Systems"; Encyclopedia of Human Computer Interaction; 2004.

[10] Wald, M. and Bain, K; "Using Automatic Speech Recognition to Assist Communication and Learning"; Procs. Of the 11th International Conference on Human-Computer Interaction; Las Vegas; 2005.

[11] Wald, M., Bain, K. and Basson, S. H.; "Speech Recognition in University Classrooms"; *Proceedings of the Fifth International ACM SIGCAPH Conference on Assistive Technologies,* pp. 192-196; 2002.

[12] Liberated Learning Project; The Atlantic Centre of Research, Access, and Support for Students with Disabilities, Saint Mary's University; 2007.

[13] Laura Schindler, Hala ElAarag; "A Speech Recognition and Synthesis Tool"; ACM Southeast Regional Conference; Proceedings of the 44th annual Southeast Regional Conference, Session: Artificial Intelligence, pp.45 – 49; 2006.

[14] M. F. Gales; "Model-Based Techniques for Noise Robust Speech Recognition"; Ph.D. Thesis; Engineering Department, Cambridge University; 1995.

[15] Michael F. Tear; "Spoken dialogue technology: enabling the conversational user interface" ACM Computing Surveys, Volume 34, pp. 90-169; 2002.

[16] Ted Maher; "History of Speech & Voice Recognition and Transcription Software"; Dragon Systems, Centaur Academic Media DBA Centaur Multimedia; 2003.

Parallel Computing

Daniel S. Priece

Villanova University
Computer Science Department
Villanova, PA 19085
daniel.priece@villanova.edu

Abstract

The rapid rate at which technology is evolving has led to a need for more sophisticated techniques that can take full advantage of the hardware available. Traditionally, software has lagged behind hardware and there is a continuous need to maximize software performance on multiprocessor systems. This paper reviews the fundamentals of parallel computing with a focus on implicit, or automatic, parallelization. We discuss different software techniques as well as restrictions that arise when creating parallel code.

1. Introduction

Parallel computing involves the use of several processes working together on a single set of code as the same time [1]. The purpose of exploring parallel computing is to cut down on the execution time of processor intense code by distributing the work among several processors. Sets of instructions are run simultaneously and then synchronized back together before producing output, instead of running serially one after the other.

The very first example of parallel computing was found on a tablet dated back to 100 B.C. [2]. The tablet consisted of three calculating areas where the user was able to compute at higher speeds. It has been an idea since the earliest digital computers were built in the 40's and developed multi-processor machines were developed in the late 60's. Today the most powerful computer is able to use 20 teraflops of processing power [1]. Rapid hardware advance has created a need to write faster and better performing software. The code that one machine could execute in 8 hours could be executed on a parallel system of 8 processors in one hour greatly reducing the actual run time, with each processor running a smaller section of the input code. Each processor is doing the same amount of work as the single processor machine but they each processor could be running a different section of the code at a time.

2. Parallel Hardware Setups

Today there are four standard setups for parallel systems differing in how the memory and communication between processors are handled [1].

The first is a shared-memory multiprocessor setup which is found in a standard multi-processor machine. The machine uses multiple processors for computing while sharing one common memory area. There is a single set of memory addresses that every

processor knows so that no copying and moving of data is required. This method is easier for programmers but because the data can be updated by one processor, if another processor is not aware of the data change it could cause problems with output.

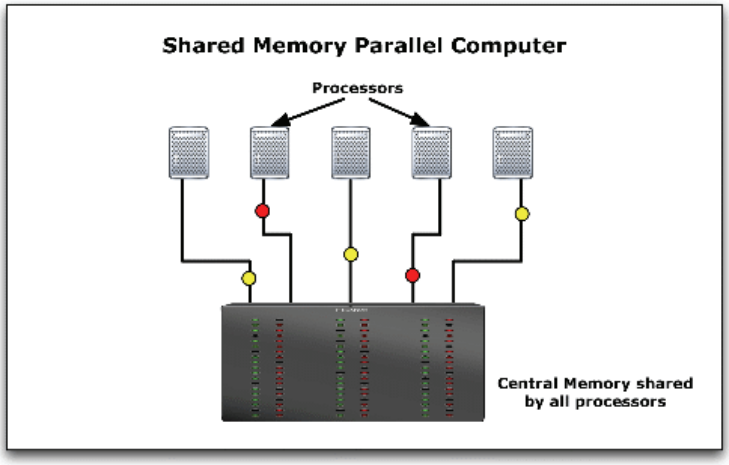


Figure 1: Shared Memory Multi-processor [14]

The next most commonly thought of parallel system is network computing. Network computing is more commonly referred to as a large network system usually operating over an internet connection. The problem with network computing in this fashion is that different computers have different processor and internet setups thus making the computations both inefficient and slow. A slow connection between computers creates a problem for an efficient parallel system. Network computing is also referred to as a distributed memory, multi-computer system.

A smaller computer network, known as cluster computers, work over a local area network connection and although they are faster than a larger internet based network, they are still limited by their communication speeds. The basic idea is still the same and is easier to manage than network computing due to the smaller number of computers in the system and fast, efficient connections between cluster computers.

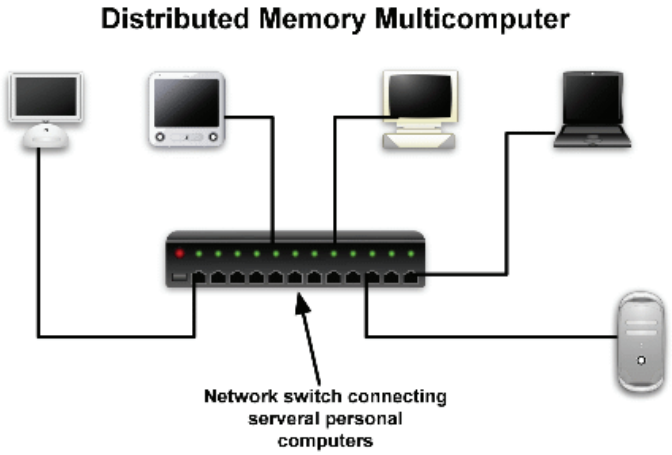


Figure 2: Network Computing [14]

The final type of parallel system is a parallel vector processor. A vector processor is a type of CPU that is able to run simultaneous operations [3]. A standard serial processor is known as a scalar processor. Instead of processing just instructions in its pipeline, a vector computer processes the data as well as the instructions. For example, if a scalar processor is used to add a group of ten numbers to another group of ten numbers, a loop would be used to add each number to the next number then add that number to the next and so on until all the numbers have been added. Instead of taking in only two numbers at a time, a vector processor grabs the first set of numbers and the second set of numbers and adds them all in a fraction of the time it takes a scalar processor. Vector processors were used for supercomputers in the past but due to high cost, they are used less today. The choice of a parallel system is important as well as the architecture of the processor and memory.

2.1 Flynn's Taxonomy

Flynn's Taxonomy provides four possible architectures for parallel computers in terms of processors and memory [4]. A standard non-parallel computer is referred to as a single instruction, single data machine because during any one clock cycle there is one instruction stream being executed using one data stream as input. Single instruction, multiple data refers to a machine that for one clock cycle, every processor of the machine executes the same code with separate inputs.

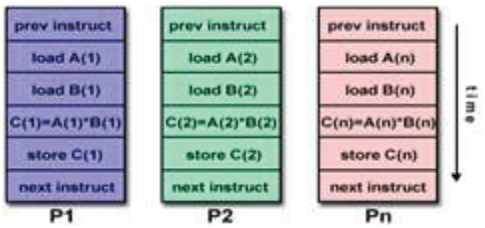


Figure 3: Single Instruction, Multiple Data Execution [4]

For example if during the first clock cycle processor 1 is executing a store instruction, then every other processor is at the same time executing a store instruction but each with different data to store. As you can see in the figure above, every processor is executing the same set of instructions however, P1 is using "1" as input, P2 is using "2" as input and P3 is using "n" as input. In an ideal situation, the processors will compute the inputs simultaneously such that they end execution at the same time and can synchronize the output. The remaining two machines are the multiple instruction, single data machines and the multiple instruction, multiple data machines which use separate instructions for each processor given a single set of data or multiple sets of data. Most modern parallel computers fall into the multiple instruction, multiple data category.

3. Programming Models

Currently there are many programming models being used, but only four of the more commonly used and less difficult models are discussed here [4]. These models are simply abstract ideas to be implemented on different hardware architectures. A shared memory model requires semaphores are used to control when a processor has access to the shared memory. This simplifies the execution in that data does not have to be sent between processors because it is all stored in a common space. Next is the threads model which allows code to have multiple execution paths by using threads to act like subroutines in a serial program. Third is the message passing model, where each process in the code has its own memory that can send and receive data from other processes. Finally, the last process is the data parallel model where processes all work on the same set of data but different partitions within that set of data.

4. Automatic Parallelism

Automatic parallelism, or implicit parallelism, involves parallelizing a program at compile time instead of implemented as in manual parallelism. An example of serial code converted to parallel code can be found in Figures 4 and 5 below.

```
npoints = 10000
circle_count = 0
do j = 1,npoints
  generate 2 random numbers
between 0 and 1
  xcoordinate = random1 ;
ycoordinate = random2
  if (xcoordinate,
ycoordinate) inside circle
  then circle_count =
circle_count + 1
end do
PI =
4.0*circle_count/npoints
```

Figure 4: Pseudo Serial Code for Computing PI [13]

```
npoints = 10000
circle_count = 0
p = number of processors
num = npoints/p
find out if I am MASTER or WORKER
do j = 1,num
  generate 2 random numbers
between 0 and 1
  xcoordinate = random1 ;
ycoordinate = random2
  if (xcoordinate, ycoordinate)
inside circle
  then circle_count =
circle_count + 1
end do
if I am MASTER
  receive from WORKER their
circle_counts
  compute PI (use MASTER and
WORKER calculations)
else if I am WORKER
send to MASTER circle_count
endif
```

Figure 5: Pseudo Parallel Code for Computing PI[13]

The set of pseudo-code above shows an algorithm for approximating PI through serial and parallel code. In figure 2, after the program inscribes a circle in a square, the serial algorithm finds the number of points in the square that are also in the circle then it divides the number of points in the circle by the total points in the square, multiplies by 4 and gets an approximation of PI. The parallel code is marked in red where changes have been made to make it parallel. In the execution of the parallel code, a master processor is setup that takes in all the information of the other processors. The worker processors each take a different section of the loop that needs computing, they each work on their share simultaneously and then send the information back to the master processor. This takes less time because more computations are being processed per clock cycle.

Because manual parallelism has many time and budget restraints, this paper explores automatic parallelism more deeply. The automation of the compiler provides a less error-prone, less time consuming, and less complex way to parallelize code.

4.1 Compilers

To understand how a parallel compiler works we will first look at a standard serial compiler which commonly consists of 7 parts [5].

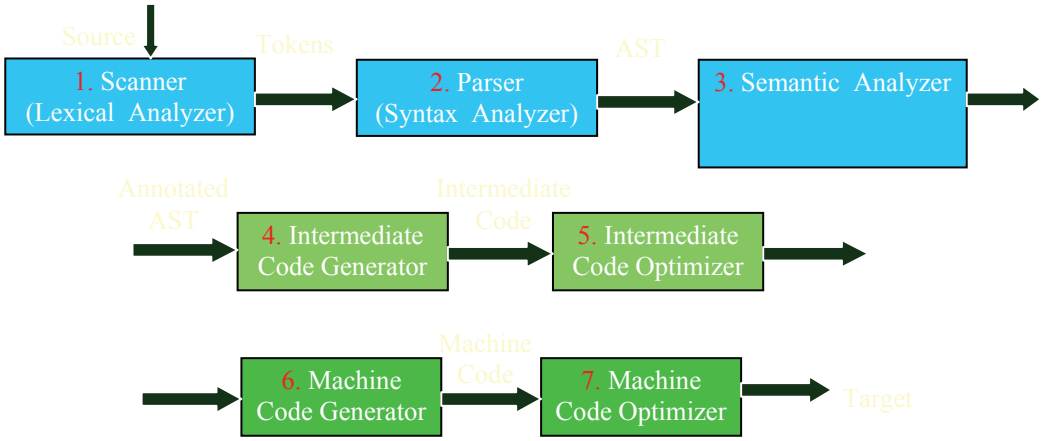


Figure 6: Compiler Diagram [5]

The first part of the compiler is the Scanner which is also known as the Lexical Analyzer. Here is where the source code is tokenized. White space is removed as well as comments and lexical errors are handled at this step. Next is the Parser. The Parser makes sure the syntax of the language is followed according to the language's context free grammar. This is where the code is tested for parallelism and separated in a parallel compiler. Then, the Semantic Analyzer handles type checking. From there an abstract syntax tree is used to generate and optimize a lower level code which is then generated and optimized into machine code.

4.2 Problems to Consider

When designing any kind of parallel compiler a few problems must be considered [6]. Latency is the time it takes to send a zero length message from one processor to another. High latency will require more planning in sending messages. Because of latency, smaller sized messages should be clustered together before sending in order to be more efficient. Bandwidth refers to the speed at which data is transferred between processors during communication. Lower bandwidth would call for as little communication as possible to be optimal. Determining which tasks have to communicate to each other is a problem called scope. All communication has to be synchronized as well which means synchronization points need to be created to optimize all the processing power and not have any processes waiting for others to finish. One major problem to consider when creating parallel code is dependencies.

Dependencies within code are the largest problems for parallelism and therefore to look at dependencies within code we use a data dependency graph.

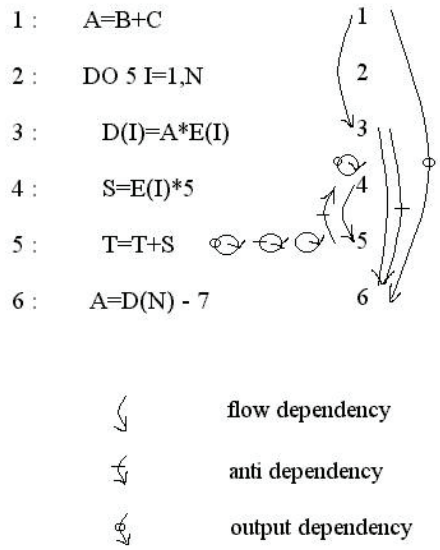


Figure 7: Dependency Graph [7]

Flow dependencies are ones where there is an execution path between two statements. In a flow dependency, a value is written and then read. An output dependency is where the same memory location is being accessed and written to more than once. Anti dependency is when a value of a variable is read and then written [8]. From a dependency graph we could optimize the dependencies so that the code can be as parallel as possible. With the remaining dependencies it is necessary to use synchronization points to optimize parallelism.

There are a few methods that have been implemented to get around the problem of dependencies [9]. Some values of a program are not known until run-time which makes splitting code difficult. For small network setups, multiple path allocation is a good method to use. It involves scheduling every possible path that the data could take at compile-time and then run the appropriate one at run-time. This obviously becomes very inefficient for larger numbers of processors because as processors are added, path allocation greatly increases. Another technique is to use a single data stream instead of multiple streams between processors. This is called dynamic destination selection. The single stream carries all of the data which is marked by the destination that the data is going to. This way the run-time systems are able to quickly send data that is dependent to the same place.

5. Related Topics

There are many compilers, systems, and tools already available that use parallel computing such as the popular systems and tool, BOINC and TPL.

5.1 BOINC

According to their website, “the [Berkeley Open Infrastructure for Network Computing] is a software platform for volunteer computing and desktop grid computing” [10]. This meaning that users volunteer their processor’s computing power over an anonymous connection or a small private network. Projects have been developed like SETI@home where anyone can sign up and use their home computer’s down time to do large amounts of parallel processing helping the search for extraterrestrial intelligence.

5.2 TPL

There are many tools available to programmers such as Microsoft’s Task Parallel Library [11] which provides methods for parallelism. Microsoft shows the efficiency of the TPL Parallel.For method for loops versus the .NET ThreadPool class where as the number of processors increase, the efficiency of the TPL increases.

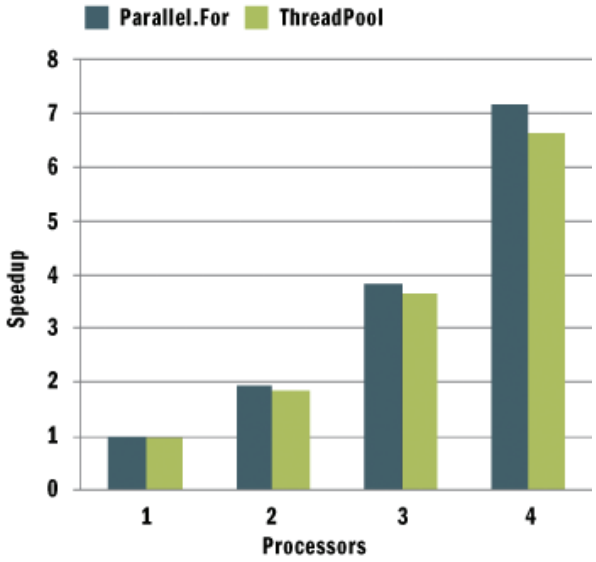


Figure 8: Efficiency Graph [11]

One example that Microsoft gives for their TPL uses ray tracing. Ray tracing is a way to produce photorealistic renderings which requires much computation [11]. In ray tracing, a path is traced from what an imaginary eye would look at and the lighting is affected accordingly. Each path of pixels is called a ray.

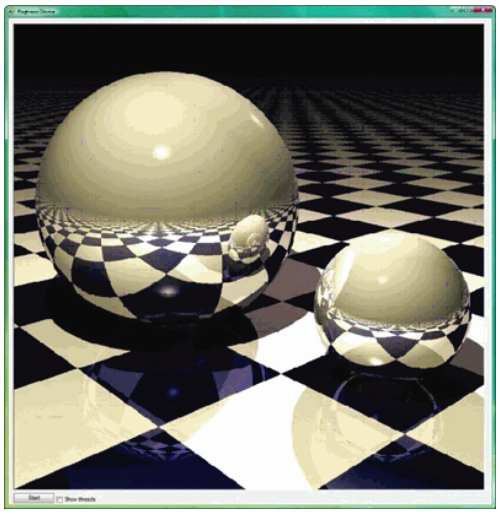


Figure 9: Rendered Image [11]

Figure 5 is an image from a rendering using the ray tracing method and the Microsoft TPL. The result was a seven times speed up on an eight processor machine from 1.7 frames per second to 12 frames per second in its animations.

6. Proposed Work

Since compilers are written to optimize code for the type of processor, it is evident that despite much of parallel optimization being done at a high level, low level optimization is more effective. Optimization of parallel compilers has reached the point where the number of processors does not matter to the compiler rather it is important at run time [15]. If there existed a compiler that could parallelize code at a lower level, then optimization across different processor types could be increased dramatically.

Dependency graphs are generated in the parser stage of a compiler. Intermediate code then optimizes the newly parallelized code and generates machine code. However, if an intermediate code generation could be used between different processor types that produces parallel code, the final machine code could be more efficiently generated parallel code because the generation and optimization is taking place at a lower level.

To study this more, we would have to study high level and low level code and look at intermediate code generated by different compilers that are optimized for different processor types. From this we could determine the areas where parallel code can be generated easily. Then developing a compiler that generates the lower level parallel code, we can test the efficiency of the parallel code produced to that produced of current parallel compilers.

7. Conclusion

In starting this research, it quickly became apparent that parallel computing is a very broad field and narrowing down on a single problem would not be very simple. The majority of information that could be found was very abstract and mostly served as an introduction into the topic. In order to understand the problem that was being researched, a very extensive background knowledge had to have been developed. It was not until starting to search for information on implicit parallelism and compilers did there appear more technical sources of information. Eventually, the topic of dependencies was brought up. This is a large problem in parallelism and ways around it must be found in order to use parallelism to its full potential. In the studying of dependencies we can understand a better way to parallelize code at lower levels thus resulting in more efficient parallel code. Automatic parallelism is a shot at better future standards of programming. If there could be one standard compiler that worked over a multi-platform system and was efficient due to low level parallel code generation, manual parallelism would not be needed and the ease of understanding parallel code would be increased. With slowing hardware development, parallel software will become more popular as well as parallel systems. With a better development of automatic parallelism, new software would be simple and inexpensive to write.

References:

[1] M. Katz, G. Bruno, "Guide to Parallel Computing," CTBP & SDSC Education Module. [Online]. Available: http://ctbp.ucsd.edu/pc/html/intro1.html. [Accessed: Nov. 4, 2007].

[2] http://www.buyya.com/microkernel/chap1.pdf [Accessed: Nov. 2, 2007].

[3] "Vector Processing," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Vector_processor. [Accessed: Nov. 4, 2007].

[4] B. Barney, "Introduction to Parallel Computing," *Introduction to Parallel Computing*, Livermore Computing, June 2007. [Online]. Available: http://www.llnl.gov/computing/tutorials/parallel_comp. [Accessed: Sept. 11, 2007].

[5] V. Gehlot, "CSC8310 Linguistics of Programming Languages," Villanova University. [Online]. Available: http://www.csc.villanova.edu/~gehlot/8310/lec/Week2.ppt. [Accessed: Oct. 28, 2007].

[6] S. Chang, "Parallelization of Codes on the SGI Origins," NASA Ames Research Center. [Online]. Available: http://people.nas.nasa.gov/~schang/origin_parallel.html#fraction [Accessed: Sept 26, 2007].

[7] "Parallel Compiler Tutorial," Tutorial-Reports.com. [Online]. Available: http://www.tutorial-reports.com/computer-science/parallel-compiler/tutorial.php. [Accessed Oct. 28, 2007].

[8] "Data Dependency," The Everything Development Company. [Online]. Available: http://everything2.com/index.pl?node=data%20dependency. [Accessed: Oct 28, 2007].

[9] C. Metcalf, "Approaches to Data Dependency,". [Online]. Available: http://home.comcast.net/~cdmetcalf/papers/cop/node67.html. [Accessed: Nov 2, 2007].

[10] "BOINC," Berkeley Open Infrastructure for Network Computing. [Online]. Available: http://boinc.berkeley.edu/. [Accessed: Sept. 28, 2007].

[11] D Leijen and J Hall, "Optimize Managed Code For Multi-Core Machines," The Microsoft Journal for Developers, 2007. [Online]. Available: http://msdn.microsoft.com/msdnmag/issues/07/10/Futures/default.aspx [Accessed: Sept. 28, 2007].

[12] "Ray Tracing," Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Ray_tracing . [Accessed: Nov. 4, 2007].

[13] "Introduction to Parallel Programming," Maui High Performance Computing Center. [Online]. Available: http://www.mhpcc.edu/training/workshop/parallel_intro/MAIN.html. [Accessed: Nov 6, 2007].

[14] "Parallel Computing." CFD Online. [Online]. Available: http://www.cfd-online.com/Wiki/Parallel_computing. [Accessed: Nov 19, 2007].

[15] M Lam, "Locality Optimizations for Parallel Machines," The Stanford SUIF Compiler Group, 1994. [Online]. Available: http://suif.stanford.edu/papers/lam94.ps

Guard Coverage Localization

Minh Kinh C. Tran
Computing Research Topics
Department of Computer Sciences
Villanova University, Villanova, PA 19085
minhkinh.tran@villanova.edu.
Fall 2007

Abstract

Guard Coverage Localization uses transmitters to help gauge the location of a particular node in a specific determined area. It is preferable to other localization methods because of the low cost and high reliability. Guard Coverage Localization lends itself to be robust. It can handle both static and dynamically changing areas. Lending itself to be of practical in many military and civilian applications.

1. Introduction

Guard Coverage defines an area P, which restricts the user's access to a particular network. To define an area, Guard Coverage uses overlapping transmitters with the ability to broadcast keys within an angular cone region. Depending on the overlapping transmissions of several transmitters, a client location can be determined from the keys the clients use.

1.1 Guard Placement

Angular Guard Placement deals seek to define an area. The placement of the transmitter greatly affects the quality of coverage over the area.

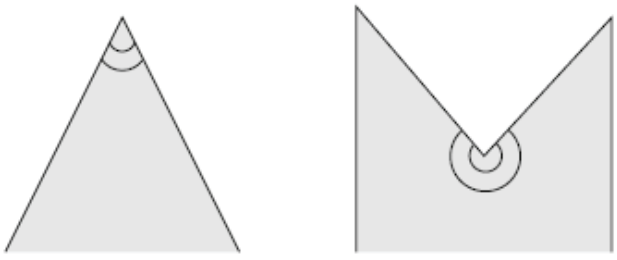


Figure 1: Broadcast angle could be convex (left) or reflex (right).

A transmitter can be placed on either a vertex of the input polygon (area) or an arbitrary point in the plane [3]. The coverage of the guard is defined by the broadcast angle of the guard, which can be set to any value. Figure 1 shows how a guard can be used to cover the interior of a polygonal region containing reflex vertices (i.e, vertices at which the internal angle

exceeds 180 degrees.) A vertex guard whore broadcast angle is restricted to the angle of the vertex it is position at is called natural. Not every polygonal area can be covered by placing natural guards at vertices.

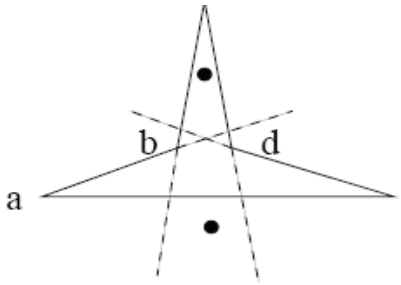


Figure 2: Simple polygonal region that cannot be covered by natural guards [3].

Such an example is illustrated in Figure 2. The two points in the figure cannot be distinguished from each other. The topmost transmitter covers an area that contains both points. Both of these points will receive the same key that the transmitter broadcast. The transmitters at a and its symmetric vertex do not cover either of the two points, and the transmitter at b and d cover both points. Since both of the points will receive the same set of keys {b,d and the top most transmitter's key} the system cannot distinguish between the two points. A 180-degress guard placed at a and facing up would separate the two points and would therefore resolve this ambiguity.

2. Guard Coverage Problems

The two types of guard coverage are used to define a certain area: are Deterministic and Stochastic Coverage [1]. Deterministic coverage handles static areas only [1],where guards are placed in specific but permanent positions. Alternatively, Stochastic Coverage can handle randomly-shaped areas that vary in time, while still maintaining their coverage [1]. Clearly the Stochastic Coverage problem is significantly harder than Deterministic Coverage Problem.

2.1 Art Gallery Problems

The Art Gallery Problem is a simplified version of Guard Coverage to whose goal is to simply maintain surveillance on the items within the gallery [3]. It uses basically the same principles as the 2 Guard at an Angle. The Art Gallery Problem uses the guards and overlaps their transmission (field of vision) to help maintain surveillance. The main issue is trying to figure out the minimal number of guards needed to keep watch over a certain area [5]. Other methods of placing the guards helps reduce the number of transmitters needed in defining an area. If there are n sides, a lower bound for the number of transmitters required is n/2 transmitters [3]. Another approach is to associate a weight with each node representing the importance of guarding that particular node. This approach is important when the number of available guards is limited. Not all the items needed to be watched that carefully, higher weighted items could have more guards [5].

2.2 Voronoi Diagram

The Voronoi Diagram for a given set of points is a planar graph whose edges partition the plane into several zone, each of which contains one single input point. Among many other applications, Voronoi Diagram proves useful in finding the Capacity of a Wireless Network [2]. This is extremely useful in communication, due to the fact that knowledge of network capacity can help reduce interference. One way to reduce interference is to prevent Signal Flooding. Signal Flooding when each transmitter communicates with all other transmitters within reach by trying to cover the most area. The Voronoi Diagram helps find the Nearest Neighbor for each of the transmitters so that they can communicate more effectively [2]. Each transmitter uses the Voronoi Diagram to learn the particular range it needs to cover in order to maintain network connectivity and effectively communicate with a subset of its neigbors.

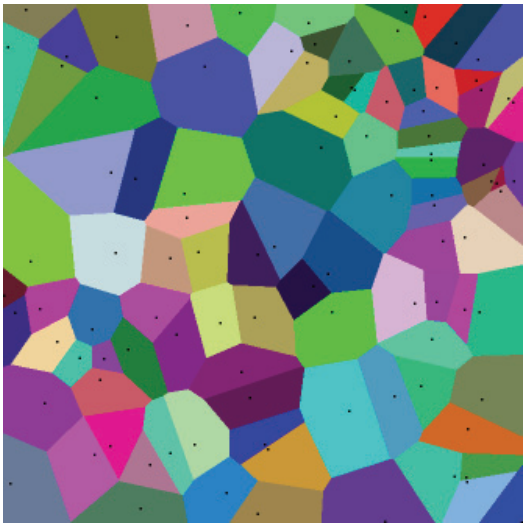


Figure 3: A Voronoi Diagram Example

In the Voronoi Diagram from Figure 3, each point represents a transmitter and is located in a object called a Voronoi Cell. The size of the cell depends on the range of the transmitter. This reduces a lot of signal flooding problems that may occur in a Stochastic coverage.

2.3 Optimal Location Problems

2.3.1 2-Guard Problem

The 2 Guard problem states that Guard1 and Guard2 can see node p if and only if the angle G1pG2 is in the range (a, pi -a) [1]. Where a is a particular node. The placement of these guards is important. The angle that the 2 guard makes defines the area covered. This type of coverage does not lend itself to being all too precise. This problem can be solved in polynomial time. The 2-Guard Problem is the most basic form of coverage to determine a polygonal region. This concept can be extended to more complex problems like the Art Gallery problem.

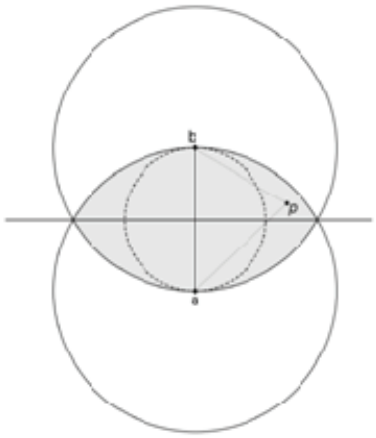


Figure 4: Example of 2-Guard Problem

2.3.2 Triangle Guarding

Triangle Guarding requires the presence of 3 Guards (Transmitters). Assuming that the coordinates of each transmitter and the pairwise distance between transmitters are known, the problem asks to deduce the length of each side and the area that it covers. Triangle guarding is more robust than using the 2-Guard Problem [1]. Using 3 Transmitters helps bound an area much more distinctly. To increase the efficiency and accuracy of Triangle Guarding one must eliminate redundant nodes. The fewer guards that the Triangle Guarding uses to define an area, the more precise the results will be [1].

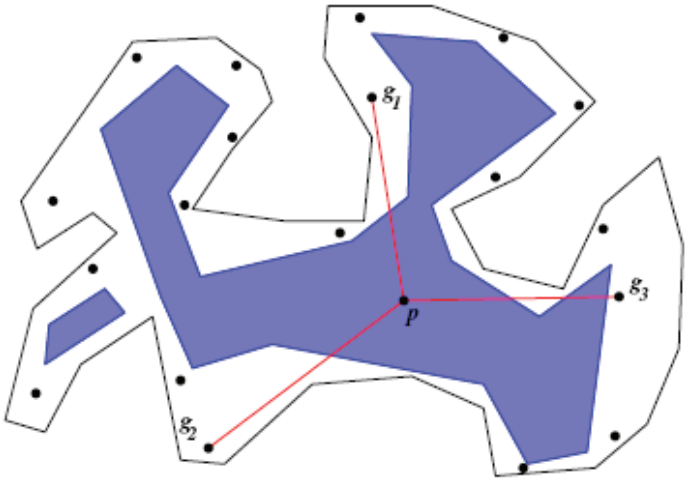


Figure 5: Example of Triangle Guarding

The diagram illustrated in Figure 5 shows an example of Triangle Guarding. S1,S2, and S3 are all transmitters, and point p is the node that they are using the Triangle Guarding on [1]. Using the coordinates of S1, S2, and S3, Triangle Guarding tries to calculate the range of area that these transmitters cover. Current solutions are computationally

expensive for randomly placed transmitters. In order to minimize the amount of error that occurs, stationary transmitters are needed.

2.3.3 Minimum Line Covering Problem

Minimum Line Covering uses a collection of lines in the plane and tries to determine whether the set of transmitters placed at endpoints of these lines suffice to guard a particular nod using Triangular Coverage.

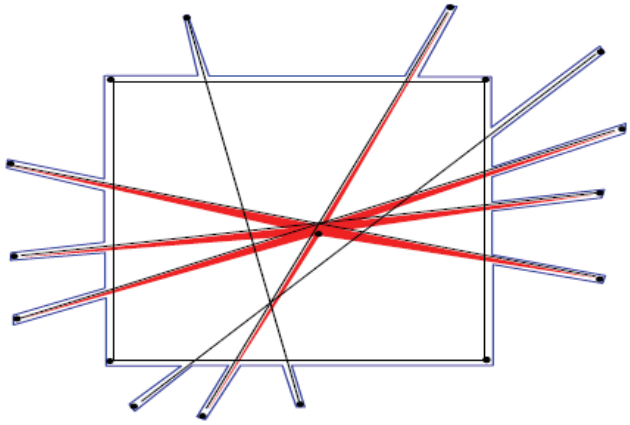


Figure 6: Example of Minimum Line Covering

The black dot in the middle is the node Triangle Guarding is used on. The other outlying nodes are transmitters that are in plain sight of the node. Each line represents a situation where 2 transmitters can cover a certain node [1]. A selection of the best choice of intersection nodes can be made so that triangulation would be more efficient [1].

3. Guard Localization versus GPS

3.1 Drawbacks of GPS

There are some drawbacks with the use of GPS [4]. First, it is very expensive to implement (in comparison to the Guard Coverage Localization). A satellite must be placed and maintained in an orbit around the earth. Repairs done to the satellite themselves take a lot of resources to do. The data received by the GPS receiver is not the most up-to-date information. The time it takes to get to the receiver, creates a precision error. Clock errors can sometimes occur during the transmission. As the signals enter into the atmosphere they can get distorted and weakened. Transmissions from the satellite might fade or cut unexpectedly and causing loss of service [4]. This is extremely harmful to the end user that depends on the service to get from one point to another.

3.2 Advantages of Localizations using Guard Coverage

Guard Coverage is less expensive to implement [4]. All that is needed is the transmitters to help define the area needed. It does not provide actual coordinates of a node as GPS would do, however it is preferable for cases where only inside/outside status of a node is required. In fact, Guard Coverage provides more accurate information than GPS. The data feed from

Guard Coverage is a lot newer than GPS [4]. This has many applications in battle field like conditions, where it would limit the transmission of sensitive information to only designated regions. This provides an encapsulation on the area and determines the field where the data can be accessed, providing a very secure environment. It is mobile, and can be set up relatively faster than with GPS. In the end, Guard coverage lets individuals set up quick and responsive networks that can help securely monitor and grant access to various nodes.

4. Proposed Future Work

We propose to implement existing guard coverage algorithms on the Lego Mindstorms robots available in Computer Science Lab in Mendel 156. These robots are equipped with Bluetooth wireless interfaces and can broadcast information within an angular range, which is consistent with the requirements of guard coverage algorithms. The developed software will serve as an educational tool to demonstrate the use of embedded systems (robots, in our case) in enforcing security within a predefined area.

To implement this project we would first have to review the capabilities of the Lego Mindstorms NXT, the capabilities we would design and construct a robot that can roam an area that is defined by Guard Coverage transmitters. This robot will also relay the transmitted keys it receives from surrounding transmitters. The transmitted keys will indicate which transmitters were guarding that particular region [4]. The transmitters, that were used to guard a particular region, could be represented by stationary or mobile robots (according to Deterministic or Stochastic style of coverage) [3]. Once the robot has been assembled the next step would be to implement Guard Coverage algorithms on the NXT. Once a stable version has been achieved we will then release a tutorial describing the process as well as software on our individual Villanova Computer Science Website. The primary goal for this particular project is to create a physical representation of the Guard Coverage Problem so that it can be used as an educational tool to help demonstrate wireless guard coverage algorithms.

5. Conclusion

The more transmitters an area has, the least accurate the reading is. Many open problems seek to find unique ways to use fewer transmitters. Associating weights with the nodes to be guarded is one of the ways researchers have tried to reduce the amount of transmitters needed to cover a certain area. In the Art Gallery Problem, items that are worth more had higher concentration of transmitters that were in line of site, where as items that were least important was no always given a transmitter straight view. This approach, however, does not work for situations in which the inside/outside location of nodes is needed besides coverage, case in which the coverage problem becomes much more complicated.

References:

[1] Meguerdichian, Seapahn, Koushanfar, Farinaz, Potkonjak, Miodrag, and Srivastava, Mani B. Coverage problems in wireless ad-hoc sensor networks.IEEE INFOCOM 2001 20th Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings, 2001, pp.1380-1387.

[2] Aurenhammer, Franz.Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Computing Surveys 23, pp. 345-405, 1991.

[3] D. Eppstein, M. Goodrich, N. Sitchinava, "Guard Placement of Wireless Localization," 14 Mar. 2006; http://arxiv.org/PS_cache/cs/pdf/0603/0603057v1.pdf.

[4] Tseng, Yu-Chee, Wu, Shih-Lin, Liao, Wen-Hwa, and Chao, Chih-Min.Location awareness in ad hoc wireless mobile networks.Computer, Jun 2001, pp 46-52.

[5] Fragoudakis, Christodoulos, Markou, Euripides, and Zachos, Stathis.Maximizing the guarded boundary of an Art Gallery is APX-completeComputational Geometry: Theory and Applications Volume 38, Issue 3.pp.170-180, October 2007