Theoretical Models and Applications: Log Book

Jordan Ell V00660306
University of Victoria, Victoria, British Columbia, Canada
jell@uvic.ca

I. May 2nd 2013

The main topic of this class was to discuss how to identify high impact papers. Some of the ideas that members of the class had to identify high impact papers are as follows. Looking at the number of citations a paper has is a good indication of how well the paper is known in its own field or how important it might be. Another is to look for prize winning papers. This can be large prizes such as the Turing award or even the best paper of a conference. This is a good way to identify a paper that has high impact potential as it is important at the time of the award. Another good way to find a high impact paper is to look at industry impact. Here if a paper involves the creation of a real world product which generates revenue it may be a high impact paper or if it can change existing products in the real world.

I had an idea for identifying central papers but did not mention it in class. My idea is a similar idea to that of industry impact and it is media impact. If a simple Google search can turn up a lot of results that all point to a single paper that has achieved main media success, it may be a high impact paper. This ties in to industry impact as the media may find out about it through industry experience. I plan on trying this method in the next assignment.

Another main area that was described this class was the discussion of what can you identify as a theory paper or a paper that has theory at its roots. It seems to me that theory is a pretty general term and as long as there is some underlying theoretical model to it. This can be some type of NP-Complete issue all the way to graph models.

The assignment for next class is to find 1 or 2 high impact papers with an underlying theory background. This is due May 6th.

II. May 4th 2013

Today I looked up my two high impact papers for class on May 6th. I took the approach mentioned earlier in my log book. That is to say that I went for media and industry impact rather than academic impact of the paper. As my interest for this course was to do with networks, I started my search on Google looking for network impacting papers.

To do this, I used Google to search for 'Coded TCP'. I knew in advanced that encoding protocol packets can be used to speed up internet connections both wired and wireless, and I knew that TCP is the highest used protocol for consumer use. This search landed me on many blog and news pages that involved a group of researchers at MIT. The group found a way of using linear equations to solve for packets that are lost in transmission instead of requesting that the packets get resent. This immediately grabbed my attention as it contains a theory background and has obviously been large enough to attract media.

I found that this new method has been implemented at MIT and has increased their internal networks speeds 16x. The paper that found this method is relatively new (2011), so it has little citations (44). However, I believe this paper can play a major role in the industry as the result of network speed increase requires no physical upgrade to existing networks. I have identified this paper, and its continuation paper, as high impact.

III. May 6th 2013

Recap from last class: Look for highly cited paper, award winning papers, ask your supervisor, and so forth when looking for high impact theory papers.

Theory papers: A paper to find when bugs get introduced into a project. A paper that provides a new algorithm that determines how well a drug molecule will work. The first two papers have very high number of citations. The chemistry papers are foundational papers in their field and found through Google Scholar. My paper on network packet loss using linear equations. Another paper on projected plains was found using survey papers which is a potential new way of finding high impact papers. High citations seems to be the most used criteria for high impact papers so far. A paper on approximations to NP-Complete problems and how they can be used. It is a Godell prize winning paper. Dijkstra prize winning paper was also said. A paper on performance analysis of networks. This paper was found by looking at industry impact. A paper on music retrieval says that the award winners in this field were not high impact even though they won. This paper has a high number of authors and also cites several other papers. A paper involving k-nearest neighbors for image searching. A paper that is central to social interactions in software engineering is now used in almost all fields of software engineering that involve communication. Found from experience, but also has 500 citations. Another paper on indexing file structure to create clusters which is the theory component. This paper again had lots of citations. A paper on dynamic programming from 1956 with a citation count of more than 11,000! Very old papers with high citation count are extremely foundational in their field.

It seems like most people today went for high citations over anything for identifying high impact papers. I personally thing for a paper to be high impact it has to have two components.

First is has to lay foundations for future research in a broad sense. For example, the first paper to explain the grand unified theory in physics allowed for a whole field of cosmology in the field. Yes it does have high citations but it also allows for a large amount of future work while also answering current questions. Secondly, I think high impact papers should have real world implications. This could be a paper on vacuum tubes which led to the first vacuum tube computer. This is a real impact that can help humanity.

We broke into groups and started to discuss our papers. My team has 3 students in it but only had 2 papers for today. The first paper was about dynamic programming and speech recognition software. This paper has over 11,000 citations! This paper lays foundational work that is being used in industry today with Apple and Google through their speech recognition software. The second paper was mine and it was about packet loss in TCP connections. This papers solves the problem of packet loss by using linear equations to solve for missing packets. We are choosing to use the dynamic programming paper as it has been around longer and has a larger impact (as of now) on the software industry than the packet loss paper which is newer and still has smaller impacts on industry.

IV. May 9th 2013

Today in class we broke out into our groups to prepare presentations for next week during class time. Here, my group's first task was to select a high impact theory paper as we had yet to agree on a paper to present. My personal belief was that our group should have presented the paper with 11,000 citations that Candy brought about dynamic programming and voice recognition from 1958. I believed this paper would be better to present because it is heavily tested and used in the industry (through Google voice, Apple Siri, and others) as well as has a high academic standing with 11,000 citations. However, from a vote of 2 of 3, we selected the paper I actually brought in which is about network encoding on TCP networks with linear algebra models to fight packet loss in lossy networks.

From here on, the day was spent carefully reading our selected paper in order for us to fully understand what was being talked about and for us to make notes as to what background information we might need. We also arranged a group meeting time for May 13th so that we can come together once all knowledge had been obtained about the paper to actually create the presentation.

I have planned to take time on the upcoming weekend to dive into the theoretical models of the paper as well as preform any background reading necessary to complete my knowledge of it.

V. May 12th 2013

Today I preformed several tasks for about 2 hours during my day to better understand our group's paper that was selected. First, I re-read the paper, highlighting all the theoretical components or background information which I felt was relevant and should be known for the paper. The theoretical components consisted of mainly linear algebra as well as some definitions which the author provides for the reader. My linear algebra knowledge was mostly forgotten so I has too look up simple items such as reduced row echelon form of matrices as well as pivot positions and Gaussian elimination as these components make up most of the linear algebra theory that was presented in the paper.

Next, I had to look up quite a few new terms that I had not previously known which were in the field of networks. I had a general understanding of what the TCP protocol was but not any in depth knowledge which was required for reading this paper. Items such as a congestion window, round trip time, and TCP-Vegas (a variant of TCP) had to be researched. I did most of my research online with Wikipedia as well as YouTube for some mathematical video tutorials on certain procedures.

The most interesting thing I learned today was that the paper actually shows a new method which is a hack of TCP-Vegas' measurement of round trip time. They trip the protocol stack into counting degrees of freedom in their system of linear equations rather than actually counting packet and acknowledgment round trip time. I thought this was a great hack they put forward on an existing system.

Through my research today, I made a couple pages worth of notes of what I deemed important for the paper and understanding its content. I will present these notes to my team at the meeting we have scheduled for tomorrow.

VI. May 13th 2013

A. Class

Today, during class, we has group 1 and 3 present their research papers to the class. Group one presented a paper that outlined a dependency graph with time approach to finding which commits introduce bugs in a software repository. This paper has a very generalizable approach to fix and defect scenarios which could be applied to many fields such as the health industry and sick patients and their treatments. With a software engineering background, I know how valuable this paper is to software engineering and what a large impact is has had. Everyone uses this paper when it comes to data mining and preventing bugs. An interesting thing to note is that the improvement papers that came after this original are not as well cited even though they implement better algorithms.

The second paper was that of a data mining technique to deal with protected data. This paper evaluated how data sets can be used in the presence of protected or blanked out data. Here, trees and enumeration trees were used as a theoretical model to be able to find sufficient data mining techniques given the constraints. I found that the team did not do a great job of stressing the impact of this paper as they merely said it was the first in its field. They did not mention any industry level applications of the paper either.

B. Outside class

Today I met with my presentation group to prepare for our presentation on the coming Thursday. Since we had all already prepared and read the paper as well as any other

sources needed, this meeting went relatively quickly. We went through the paper and marked down the key points we wanted to discuss as well as any external source points we thought were relevant to the paper. We set up a skeleton of our slides and prepared the presentation.

The only problem we had today was that no one could put into words why our paper's solution to a problem was better than others. This problem was left for everyone to figure out for our next meeting on Wednesday.

VII. May 14th 2013

Today I worked solo in preparing my section of the slides for my group's presentation on Thursday. I am in charge of the beginning and end portions of the presentation. I did a bit more research as to why our paper is high impact and found lots of interesting cited by papers as well as some companies that are using this paper's ideas. I also found that some of the authors spawned their own company from the ideas that they had.

I solved the issue previously mentioned on how we could not explain why this paper's solution was better than standard ones. I found slides of the paper given at a presentation at the conference. The slides explained more in detail how their solution was better than standard TCP protocol. I will share this solution at the meeting on Wednesday.

VIII. May 15th 2013

Today our group met briefly to go over our presentation for the final time. We went through our newly created slides in order to give a brief explanation of what each of us will be talking about so there are no surprises. Everyone seemed to know exactly what they will be saying as there were no issues.

I also conveyed my answer to the question we had earlier as to how the new system is actually better than just TCP. Everyone agreed that the answer I found was correct and should be used in the presentation. We will be presenting tomorrow morning.

IX. May 16th 2013

Today there were two group presentations presented in class. The first presentation was my own group's presentation. As per our assigned roles I was in charge of basically the setup to the paper and the discussion to be lead afterwards. I explained the motivation and high impact factors (which I thought our group had the best of) to the class while also giving a very brief overview of the paper to the class. I lead the main parts of the discussion after the technical parts were also presented. I think our group did a very good job on the presentation and had one of the better presentations over all to the class as they seemed engaged throughout.

The second event of today was the 2nd presentation by group 4. This presentation was on the fire fighting problem. I did not like how their paper was not actually published as that defeats the point of high impact papers as well as peer review. I did like how the problem was presented in the general overview, but the technical details were a little over my head and it was difficult to follow along. I wished the presenters would have motivated the problem more and shown some real world computer science examples of the problem and how it was actually solved or how this method could have helped solve it.

X. May 19th 2013

Today I was at the Mining Software Repositories 2013 (MSR) conference in San Francisco. Here, a paper was presented where the authors talked about determining what a commit in a repository actually did to effect method signatures. They said this was accomplished by using abstract syntax trees. This got me to thinking how they went about diffing two trees as I thought that the isomorphic graph problem was NP-complete. This got me into Googling the problem and I found out that the isomorphic problem actually has a polynomial run time for particular graph structures, one of those being trees. So what they were talking about in the paper was possible. This research is right up my own interests alley because I am working on a problem right now that requires the diffing of trees as a possible solution. I did some more research and found a paper published in 1995 about abstract syntax trees of languages and how they can be used to find the result of a code change in a software repository. I also found a tool developed by some software researchers that implements the basic algorithm in this paper while providing some good user feedback on what was found in the algorithm. I will surely be using this tool and paper going forward into my own research for my masters thesis.

This type of research will lead me to graph isomorphism into the future and how one can create some sort of set of edit rules which will show how to transform one tree into another. This also leads to the question in software of when two pieces of code are really different. For example getHeightValue() and getValueHeight() might actually be considered the same. I will bring this up during the next class as a potential topic for the next round of papers.

XI. May 23rd 2013

Today I was at the International Conference for Software Engineers 2013 (ICSE) in San Francisco. Today an interesting, and winner of a distinguished paper award was presented, using some theoretical model known as entropy. Now, in my hobby life, I am very interested in astronomy and cosmology so I am very acquainted with the idea of entropy as the 2nd law in thermodynamics. The law of entropy states very simply that as time progresses, things tend to become less organized and more uniform by nature (See the big bang which was highly ordered, to the universe now which is highly unordered). The author of the presented paper used this idea to determine what type of state a file is in after multiple authors have written code into it and what types of developers will touch it going into the future and how bug prone such a file could be (highly unordered). It was a very interesting way to approach the problem and offers a more general approach to the problem of determining if a file is bug prone in software engineering.

The problem was previously solved by simply looking at the number of bug fixes that involve this file.

With this idea of entropy applied to files in software projects, I am wondering if we can apply entropy to anything else inside software as the idea of entropy can be applied to most situations in life. I feel like a good study would be to measure the entropy of a software project over its entire life span. This may show that a software project often goes from very structured and rigorous effort, to a complete sense of chaos as the project evolves over time. It would also be interesting to see if this is true for all projects or only projects that are eventually abandoned and not in those projects which are normally deemed healthy.

XII. May 27th 2013

Today during class, I presented the two (previously) talked about papers from the research conferences I had recently been at. People seemed very interested and wanted to know more about the papers. It was also told in class that our next round of high impact papers should focus on graph theory specifically relation to graph display, similarity, and isomorphism. Since the paper I found out about at the research conference deals with isomorphism and similarity I will start my search here.

After some digging today I found the 1995 paper again and some of the theory papers that it has referenced. The graph paper talks about the issue of source code meaning the same thing but looking different (a measurement of similarity). For example getHeightValue() and getValueHeight() might actually be considered the same, so we need a similarity measurement in the abstract syntax tree graph to determine this. The theory paper that solves this issue is from 1945 and is titled Measures of the Amount of Ecologic Association between Species. It has over 3500 citations. The theory paper gives a very simplistic way to measure if two items in the same context are similar, which is what ends up being used in the tree difference of the application paper. It is interesting to see the citied by papers of the theory paper as they range across a wide variety of science topics including computer science. These will be the two papers I present at the make up class this Wednesday. I will write more about this in my Wednesday log.

XIII. May 29th 2013

Today we has a make up class for last week's canceled class. I will take this time to write some quick notes on some of the more interesting papers that were presented in todays class. The first paper talked about island parsing for software engineering. It was interesting to see how natural language can be parsed in this manor in order to figure out general ideas of that is being talked about. Another paper of interest to me was about cyclomatic complexity in graphs and how this idea can be used to measure the complexity of code. I actually want to read this paper because just the other day I was having a conversation with a colleague about how we could measure a software project's complexity at every commit in its source control repository. I may have to come back to this measure. There was another paper that was very similar to mine that I presented on edit distance and determining if two objects are alike or how they are related to each other. This would be an interesting person to be in a group with as it seems like we have similar interests.

I presented the paper I talked about in last class how I discovered it in the research conference I was just at. The papers in theory and application deal with node similarity in environments or graphs. It was used in the application paper of differencing syntax trees of source code. This is right up my research alley as my masters thesis is based on this idea. I liked another paper presented about identifying communities in networks. This papers talked about social structure does not identify groups of people but rather context of conversation can be used to find people who should be grouped together. The application was based on open source software projects which is interesting because OSS tend to be slightly chaotic and self organizing in terms of their social structure. Using social structure to predict material turnouts is an interesting topic.

There was one person in the class who failed at finding a high citation paper for the assignment. His story of citation count could be a lesson for us all as citation count is not always the be all end all of high impact papers. It seems like sometimes you find a very high potential niche of research but the problem is that these papers might be difficult to understand or that the topic has not caught on yet among a large group of researchers. This obsession with citation count I feel is bringing us down and limiting us to what we find interesting or high impact in our own eyes.

XIV. May 30th 2013

Today there was a regular class in which we again presented the next round of high impact papers selected by the class. This time the papers were presented with a little more depth so that everyone could understand what types of papers they were. Once the papers were presented, we decided to label the papers as topics and cluster them together according to their underlying theoretical models and applications. We came up with several categories, most of which had graphs as the underlying theoretical model. This is not very surprising in a room full of computer scientists. To me the graph model is probably the most interesting in regards to my field of research. The paper that I presented today also had a graph (tree) as its theoretical model. To me, from what I have seen in my time in software engineering research, not enough effort is put towards the theory component of the research. Researchers just decide to use graphs on a whim because they have a good visualization of the data they are using. This usually results in graphs of social networks or the way code interacts with one another. For my own research I have been trying to find how graphs (more specifically trees) can be used to aid rather than display my work. The paper I presented today is actually from my our master thesis topic and deals with the construction of trees in software and similarity measures between the trees. Before I found this theoretical model, I was simply diffing

blocks of code to see what was different about them. This new model gives me better knowledge and understanding to what the actual differences are.

Once the topics were formed, we split into groups (I am group 1 with 4 members including myself). Here our group simply organized a time that worked best for us all to meet and then said have the papers read by then and be ready to discuss them as well as create the presentation. I look forward to working with this team as they all seem motivated on the subject.

XV. June 2nd 2013

Today I started to read my group's papers for our upcoming presentation to the class in about 2 weeks time. I started by reading the theory paper. I honestly did not enjoy this paper as much as I thought I would have. The author seemed to go out of his way to make the paper seem more complicated than it really was. All the paper boiled down to was having a stepping algorithm of a finite set of grammars for a particular language. But the author's linguistic choices made it very difficult to understand. I think research, especially in theory when the terms and ideas can become very complex, is at its best when explained as simply as possible. The second paper I read was a lot better. The second paper was the application paper that our group will be presenting. Here the paper talked about a way to extract particular items from a body of free text using something called island grammar parsing. As I was reading this paper, I realized I had actually seen it before. About a year ago I had been doing some work for Dr. Daniela Damian and needed the ability to extract stack traces, file names and a host of other technical terms from a body of work. While the technique I selected for this process was not this application paper, it did reference it and use components from its design. It was interesting to see how high impact papers can find their way into even my own research.

After I read both these papers I made some quick notes on any questions I might have for the group or what I think were the key points that needed to be presented to the class. Our next meeting is scheduled for June 4th where our group will meet and set out creating our presentation.

XVI. June 4th 2013

Today was very short and sweet. Our group met, except for one member who did not show up or answer any emails, to discuss our upcoming presentation to the class on our theory and application papers. Here it was decided that we would present each paper for 15 minutes and have two presenters for each paper. I was slotted to present the second half of the theory paper. After that the group talked about general discussion points on the two papers to make sure everyone was clear on the ideas. We decided from here that everyone should be responsible for creating their own slides for the presentation and that we will meet sometime next week (before our presentation) to go over all the slides and to make sure we are ready for our presentation.

XVII. June 9th 2013

Today I started to create my presentation for our next group assignment. My task for this presentation is to present the detailed explanation of the theory paper our group has chosen on finite state cascades. I had a lot of difficulties with understanding this paper in general. First, I am fairly familiar with parsing test with grammars and using these types of syntactic rules in finite state automata. However, the author of this paper presented his ideas in a very confusing manor. Part of this may be because the domain knowledge of parsing free typed text is generally new to me, but I believe the author just did a poor job in explaining his ideas. There were lots of concepts in the paper I had to look up because the author mentioned them in a sentence and then never explained what they meant. On this subject as well, there were items that I could not look up because they were being presented as novel in the paper but very little to no explanation was given. Because of these limitations, I feel as though my presentation has suffered a bit. I has to throw out some concepts to present because I did not feel comfortable with my knowledge in them. On the positive side though, the main algorithm of the paper will be presented in full so I feel as though the losses are not too bad. The presentation will also be mostly focused on how the authors parser holds up to existing parsers in a rather large evaluations. However, this evaluation caused some issues for me.

While reading the evaluation and trying to come up with presentation material, the author compares his parser against pre-existing ones. This was a problem for me as just given the name, I did not understand how these parsers are different from the one presented in the paper. To be able to fully explain the evaluation to the class I feel as though I should look up the general ideas of these others parsers. However, this will be left to another day.

I finished my slides today and have some items left to look up for another day to fully prepare my presentation.

XVIII. June 10th 2013

Today I continued where I left off for the presentation creation for the second group assignment. As was stated in the previous log book entry, I needed to lookup some of the free text parsers that were mentioned in our paper's evaluation to fully understand what the author is comparing his results against. I was able to find most of the older parsers mentioned in the paper and got a decent understanding as to how they work in comparison. However, some of the papers were quite vague on the underlying concepts and just went straight into a detailed explanation of their particular variant of the algorithm. There is not too much I can do at this point without becoming intimately familiar with the domain knowledge of parsers and particularly free text parsers. I feel as though I have enough knowledge right now to give an abstract view of the presentation to my class mates which should be good enough.

I plan on meeting with my team sometime this week to prepare for our presentation on Thursday. I have had no contact with 2 of the members since the last meeting. Braden's slides

and my own slides are the only ones to be completed to this point. I am starting to become a little worried about the other two group members, but will give them some time yet before I ask about their work.

XIX. June 12th 2013

Today our group met to discuss our now made presentation and to make any last minute changes that would be needed if we deemed it necessary. Everyone in our group showed up except for one member. This member has been the most difficult part of this round of presentations. We have emailed him several times to no response and he has not shown up to any group meetings. Without this member we went ahead and made the full presentation anyways. Our meeting was very short as everyone was fairly aware as to what we wanted to talk about and how long each person's part of the presentation would be. The only item we really had to work on was our guiding questions and discussion section. We decided not to ask any specific questions about the algorithms presented in the paper as we have found that too technical questions result in low discussion among the class. I have noticed this is actually a common theme among all presentation so far. Class members seem interested in a topic up until the presenters put 3 or 4 slides in a row which is just a wall of math. People become disinterested and lose focus on the presentation. It should be up to us as presenters to understand the low level math behind the theory but at the same time be able to present it in a high level abstract manor that not only informs but also perhaps entertains the audience. I am going to try and incorporate this philosophy in my presentations going into the future. After this meeting our presentation was ready to go.

Today there was also a makeup class for a previously missed one due to a conference going on at UVic. I was actually unable to make the class as I had some convocation duties to attend to that afternoon. However, our missing group member showed up to this class. I was later informed that he was given some slides to present (after some disappointing conversations about how he did not respond to communication from the group.) Everything did seem to work out in the end as our group felt ready to present the next day.

XX. June 13th 2013

Today, our group presented our theory and application papers in front of the class. However, before we went, a group presented before us on community detection inside of networks and graphs. This was a very interesting presentation to me as it fit in with some of my research goals right now. I am working on a project which detects communication efficiency basically among open source developers all working on the same project. However, this project suffers from not understanding which components are present in the software being built. For example there could be a database, three back end drivers and 2 front end handlers. I mentioned in class that the algorithm presented may be useful for solving this problem without an intimate knowledge of the code base or manual inspection. One could list files as nodes in the graphs and their logical file coupling as their weighted edges. Logical file coupling occurs when files are frequently changed together. This way hopefully the communities detected will be the components that make up the software system. This is just my idea and perhaps it has actually already been done. I will have to do some more research into this area to see if this idea or at least one similar to it has been created.

Our group presented our papers with relative success in my eyes. As stated before, we tried to keep any math heavy components out of our presentation which I think really helped our presentation as well as keeping the discussion questions as high level as possible. We had some very interesting discussion points about accuracy vs speed in natural language processing (NLP). Someone brought up the idea that in certain situations, both accuracy and speed are needed such as air traffic control. Here, speed is needed to handle all the planes as well as accuracy to make sure what is being communicated is correct. As NLP moves forward going into the future, the speed vs accuracy debate will be lessened (will not be eliminated in academia) due to hardware performance and Moore's Law. There will be two more presentations on Monday that will wrap up the 2nd round of papers.

XXI. June 17th 2013

Today was a normal class with a single presentation in it. The presentation was on a theory and an application paper. The theory paper was talking about graph cuts. Graphs cuts seemed very similar to the similarity measures that had been previously described in class. Basically it boiled down to identifying how nodes in a graph relate to each other and then breaking the graph apart based on those similarity measures between nodes. It would be interested to compare the community detection algorithms presented last week to this new idea of graph cutting. Both of these tools could be used for identifying clusters of objects among background noise. My idea would be to identify components in a software project and septate them from the background noise. The files would be the nodes and you would just need some sort of similarity measure between them. I feel like this would be an interesting application of the theory used in both of these papers instead of manually inspecting a large software project.

The application paper presented was in my opinion the most interesting application as it had objectives that almost everyone in this class is familiar with, in that is it image editing. They showed a tool which could pick foreground objects out of images quite well. I would be interested to see if these algorithms are fast enough to work in real time systems such as sporting events where a broadcaster may want to track a particular player on the screen and show some sort of statistics above his head in the game. I know some broadcasters already do this but they may be just on replays or pre-rendered film or they could be using a different technology.

Finally in this class we discussed the upcoming midterm. I feel pretty confident that I will be able to answer questions to do with other papers that I did not present. I will possibly

be writing about my study questions or research in this log in the next 2 or 3 days.

XXII. June 18th and 19th 2013

I have decided to combine these two days of my log book as they pertain to the exact same activity of studying for the up coming midterm that is on Thursday. To study, I started by going over the presentations from the first round of group projects and then moving onto the second round. Here I actually had more difficulty than expected. For one, groups who posted gratuitous amounts of math on their slides were the most difficult to follow. I think as a class we should get away from posting math on our slides or going into long 5 or 10 minute talks about how every equation in our papers work. To me this is just not a good way to present the material. As a class, we should have a good enough understanding of the papers we are presenting to be able to give an abstract view of what is going on but at the same time in that abstract view provide all the details necessary. The presentations become extremely hard to follow when it is just equation after equation and people get lost or confused. So after each paper's presentation slides, I went into the paper to confirm any of the questions I still had left over. The papers in the first round were more problematic for this. The firefighting problem paper was not well written and I had to use Wikipedia to look up a lot of terms that were not explained in the slides or the presentation. Wikipedia seems to be almost better than our slides at giving quick explinations for some of our topics.

The second round of presentations went a lot better for studying so I was able to breeze through them very quickly. One thing I did notice about the second round papers is how similar the community detection algorithm and the image segmentation papers are. They almost use the exact same theoretical models and algorithms. Both papers represent their problems as graphs, then computer some sort of similarity measure between the nodes (this is where the papers differ). After that the papers use these similarity measures to cut the graph in some way or split it up into different componenets. I feel like the algorithm for one paper could easily be applied to another in this situation. I really enjoyed the community detection paper for studying as well as it gave me lots of ideas for applications in my own research at UVic.

After studying for these past two days, I feel pretty confidenet that I can answer most questions to do with any of the papers presented in our class thus far. The only concerns I still have about the midterm is how low level the questions will be. Most of the algorithms I can carry out in the papers, but the firefighting problem solution is still very difficult for me.

XXIII. June 20th 2013

This log book entry will be very short as we just wrote our midterm today. The midterm went very well for me and I think I answered all questions with relative ease. One note that I thought was interesting was that I actually answered one of the exam questions in my previous log book entry. The question was how can one paper's solution be applied to another paper's problem. I mentioned before that community detection and image segmentation are very similar and that one's problem could be solve by the other's solution.

The only other thing to note about the midterm was that I found the paper that started Google. It was the actual introductary paper for the Google search engine. I just thought that was very cool.

XXIV. June 23rd 2013

Today I did the research online required for our assignment which is due on Monday June 24th 2013. This assignment required us to pick a topic we find interesting and search online for multimedia tutorials and articles which could be useful for teaching other members of the class. I found this assignment a little difficult because the topics I am truly interested in have no real public appeal to them so the amount of multimedia presentation online are quite lacking. First I started with tree / graph edit distances. However, I really only found either wikipedia articles pertaining to isomorphism or I found academic papers of authors trying to solve minimal graph edit distances in very specific cases. People in the general public or even students at universities for the most part would have no interest in learning these topics so I found that the articles online reflected that by not having engaging learning media for thew topic. I can't tell if this is a serious problem in academia, that only the more public topics gets such attention to warrent multimedia learning, or if that is an appropriate response to the number of interested parties. After the tree edit distance problem failed (which is sad because that is the most interesting topic to me) I moved on to community detection. Here I found a plethera of tools availible online which all preform some sort of community detection and most of them offer many different algorithms for solving this issue. I thought perhaps this would be a good starting point for comparing and contrasting different community detection algorithms and learning how they work. However, the problems lies in abstraction. The tools themselves, for the most part, do not explain how the various algorithms work, they just abstract that information away to the algorithm name and preform the actions for the user. This made me look up the explinations of some of the algorithms and again, for the most part, all I found were technical research papers on how they work, besides the more known algorithms which had a couple youtube videos for how they worked. This again did not feel right to me as it seems when dealing with community detection, the solutions are abstracted away form the user in order to generate an easy to use tool. This being the case, I did not pick this as my topic.

Since I am very interested in data mining and community detection, I thought why not combine them. I decided to go with cluster analysis on graphs. Here I found an algorithm called k-means (Wikipedia[1]) which I remembered from my data mining class I took. K-means uses the Voronoi cell (Wikipedia[2]) theoretical model in graphs in order to generate

[1] http://en.wikipedia.org/wiki/K-means_algorithm
[2] http://en.wikipedia.org/wiki/Voronoi_cell

the clusters of nodes present. I found some great wikipedia articles on cluster analysis and Voronoi cells as well as some youtube videos (Youtube[3] Youtube[4] Youtube[5]) explaining all the math behind the algorithm which walks the viewer through step by step. The best part about this algorithm is that it can be applied to problems we have seen before. I found an online presentation (UofA[6]) explaining how k-means can be used to segment images into their appropriate parts. This is the same issue as seen in the second round of presentations. I think k-means could also be applied to community detection as also presented in the second round of presentations. I have decided to present k-means to the class tomorrow as my topic, along with all the media sources I found.

XXV. June 24th 2013

Today we had a normal class where all of our material we found over the previous week was presented to the class. This material covered learning resources that students found and would like to use for the last month or so of this class. There were basically three main fields that were presented. The first the foremost being Markov chains. Over 4 students presented information on Markov chains and had material for the class to use to learn. For the most part the material seemed quite interesting although it is hard to tell without getting my hands dirty on the meat of the subject matter. I plan on going over lots of the student material presented in class today in the coming weeks.

The next, and most interesting, topic presented to the class was about procedural generation, specifically in music. I personally love procedural generation, from simple L-Trees in two dimensional space all the way up to fully generated worlds such as in the video game Minecraft or in generating music. I have a great interest in this topic as I know a lot of companies such as Pixar and Dreamworks are starting to move away from artist drawn everything and more into procedural generation, especially for large scale productions. The topic presented today covered the generation of simple music, however a book was mentioned that I know from the past as to being extremely popular in computing. The book covers lots of interesting snippets of code as well as a theoretical background and builds its way up through procedural generation of music. I might go out and pick this book up as I have heard a lot of great things.

After today's events, it seems like Markov chains and procedural generation will be the key topics going into the future of the class. This is a good combination because you often find Markov chains in procedural generation.

XXVI. June 27th 2013

Today, we has a guest lecture in class that was extremely interesting. I will keep this log book entry short and just go over the key points and what my general thoughts were. The talk was all about the modeling of disease and bio-disasters such as the spread of smallpox and west nile virus. It was interesting to see how mathematical models, specifically computational models can be used to figure out how the virus will spread and what the best course of action is. This is very similar to the fire fighting problem that we learned about earlier in the course material. While the meat of the subject is interesting, the best slide I saw during the lecture was the process slide for developing a scientific model of some event or phenomena. The guest speaker had a slide that showed an iterative process where the modeler should go through a series of steps which eventually loop back on themselves for further refinement of the model to be generated. Not only does this process apply directly to modeling disease outbreak, it is also a perfect example of agile software development which I thought was an interesting connection. The fact that some analysis of the overall problem is done, followed by an iterative refinement of product is just a perfect connection between software and mathematical modeling.

Other than the stated above, the guest speaker was very engaging and had great material for our class to learn. I would hope to see more guest lectures moving into the future.

XXVII. June 30th and July 1st 2013

I am writing these two days of my log book together as they pertain to the same work items preformed by myself. Over the last two days, I have spent some time looking through some of the student suggested material for further learning in the class on June 24th 2013. I started by watching the video series presented by Braden on Markov chains. These videos, while a great introduction to the material, seemed very slow at first for me. The first three videos basically stepped through how to use matrix multiplication and iterative loops to explain and show the use of Markov chains. This was pretty tedious for myself but a good refresher non the less. The whole time these videos were running I was actually wondering to myself if Markov chains always converge with the more steps you take. Thankfully I did not need to look any further as the last 2 videos in the series answered this and many other similar questions. So overall this was a good video tutorial once you get past the beginning.

Since Markov chains are not exactly my favorite subject, from here I moved onto procedural generation of graphics. I have had some experience in this before with L-Trees but nothing too major. Aside from the material posted, I found a great website on the material(Procedural Generation[7]) and decided to go over it in details. This website has a tutorial on how random fractal terrain in generated in 3d video games. It goes over everything from the theory behind the implementation all the way up to and including a working example along with the source code of the algorithm. I really had fun reading this tutorial and it makes me want to make a small video game as a final project to this class where the world is dynamically generated using these philosophies.

[3] https://www.youtube.com/watch?v=0MQEt10e4NM
[4] https://www.youtube.com/watch?v=4shfFAArxSc
[5] https://www.youtube.com/watch?v=aiJ8II94qck
[6] www.cs.ualberta.ca/~nray1/CMPUT466_551/Clustering.ppt
[7] http://www.gameprogrammer.com/fractal.html

This is it for now. I do not know how far we are suppose to go in our own reading of this material so I will stop after these two examples listed above. Hopefully we talk more about procedural generation in class. I look forward to hearing Nick's ideas on the subject as I know he has had some experience in procedural generation in music.

XXVIII. July 2nd - 5th 2013

I am again bunching up my log book entry into a single entry as the following material is generally from the same experience over the time frame given. This week I started working on my final project for the class. Since my own research is in areas of software repository mining, I decided to stay with that trend. Something I have always been interested in is software evolution of a project over its life span. What I mean here is how the code changes in its actual meaning over time. To preform such an action, code needs to be compared with previous version inside the project. This means building abstract syntax trees of the code and then comparing them, which definitely means edit distances of trees as previously studied and mention in this log book and class.

I started by looking online to see if there was any papers on this exact material, thankfully I found a paper where the authors created a tool known as ChangeDistiller. Here they almost implemented exactly what I need. They defined create, move, and delete actions on a tree while as creating a taxonomy of change types in Java source code. There taxonomy included items such as adding a method or changing a class field. This tool they created was really great so I decided to use it going forward. I learned how their source code operates and even made some slight modifications to it to include exactly what I needed.

Once this base project was found, I went forward with creating my own project. I decided a web interface would be great for a visualization of the data to be generated. Braden helped me in creating this front end interface. After that was built, we simply defined what types of changes we cared about for the project we were studying and ran our new program on pre-existing project with these metrics. The whole development process was actually very smooth and we ended up with quite a nice little tool(API Evolution[8]).

We now have some research questions we would like to answer in terms of actually doing something useful with this information, as well as possibly publish our work. I am interested to see how the project behaves around stable release points of the project's milestones. Does the project slow down around these points? Are certain metrics high than others? The list of questions could go on and on with this. I look forward to the final paper of our project and answering these questions. As the month goes on I will provide small updates in this log book with any findings.

XXIX. July 4th 2013

Today we had a normal class. A good chunk of the class was devoted to going over the midterm. I was very pleased with my midterm grade overall so I did not pick up that much additional information here. The remainder of the class was used to move forward with some of the student material found over the last week. We moved forward mostly with Markov chains. Unfortunately we started watching the same video series that I already saw and mentioned in the previous log book entry. This is pretty much all that happened in class today. So I will keep this log book entry short and sweet.

XXX. July 8th 2013

Today was once again a normal class day. The beginning part of the class was spent talking about how each student is doing in the reading material and homework that was found in previous classes. As was stated before in my log book, I have watched the Markov chain tutorial videos as well as delved deep into the procedural generation material. Unfortunately it seems like I was the only student in class interested in procedural generation as everyone else presented on Markov chains. I would have liked to learn about this topic, however I do recognize that projects must be started and other students are not doing this topic for their projects.

After, we broke into groups to discuss various topics. Our group talked about the final projects and if there are any other ideas that our classmates can give us or suggestions for our own topic. For the most part this session went very well. Laura had some good ideas about database modeling and how to extract data in appropriate manors for a project she was doing. We tried to help her by identifying what the actual theoretical model was in her application. Eirini's project involved community detection algorithms across networks. She explained some unique data sets such as Marvel's super hero communication data set which was very interesting. I could not help her much on her topic though as I am not as familiar with community detection as she was. Lastly, myself and Braden presented our project which is actually nearing completion in terms of data gathering. We showed our tool, as previously mentioned in this log book, and got some good feedback on what types of research questions we could answer. Questions such as creating predictive models for when code stability is likely to occur in a project.

For next class I hope to bring in a list of possible research questions for our data set and get a group's opinion.

XXXI. July 11th 2013

Today ware once again a class day. This class day was spent explaining our research topics to the remainder of the class and how we will proceed with our projects going into the remainder of the semester. Myself and Braden were actually in quite a unique situation for this class as we had actually already completed the software implementation component of our project. This allowed us to show off to the class what had already been completed and collect better feedback on our potential data set as well as future research questions. The rest of the class continued by everyone showing or explaining their project in more detail than before. The projects I really got excited about were Nick's procedurally generated music,

[8] http://ballroom.segal.uvic.ca/

and Laura's implementation continuation on a stack overflow project I had seen before. I think these two projects show great promise for going into the future.

XXXII. July 15th and 18th

I decided to lump these two class days together in my log book because they were very similar in what was accomplished in each class. For these two classes, we focused our attention on a set of homework problem specifically about Markov chain models. For the first class, we broke out into small groups and attempted to solve some if any of the problems. On the first day my group was tasked with solving one of the problems (I believe it was problem number 2). And on the second day I believe we solved question number 4. My reactions to this exercise was that it involved a certain level of chaos. Our groups had little knowledge as to the subject matter of these questions and had to look up virtually everything in order to come up with any type of answer. I guess in the long run this ended up becoming a good thing as we were all forced to dive into something we had never seen before and use our skills as a group to come up with some answer. On the other hand though, it was often difficult to understand other group's explanations of the answers as we were not familiar with their question or the terminology they were using. I think this exercise is a good summary as to how most of the group work has gone in this class for me. Actually solving a problem with my group was a great experience and I feel as though we all learned something through the process. However, when other groups explained their work it was very difficult to understand what they were saying because the domain knowledge was missing entirely. This may also be attributed to students not making the best teachers or not having the proper presentations skills.

Another quick thing to note about these two class days was that students were allowed to bring in learning material for one of three topics we had previously decided on in class. The most prevalent was Markov chains which is why the homework set was done. However, myself and a few other people brought in learning material on procedural generation. I just want to write a quick thing about finding the material through my experiences. What I found online is that procedural generation is more prevalent in the arts (video games, music, graphics) than anywhere else. This being the case, it is a very popular topic in those communities. However, what I found was that while there is a lot of material out there on the subject, most of it is custom implementations of very specific generations for specific tasks. What I mean is that there is a serious lacking of generic learning materials to help students or whoever discover the background of the topic. This is probably because the background information is very abstract and does not need to be learned so that the main algorithms can be implemented. This all being the case, the material I found to bring in was mostly code examples or finished demos. The two other resources I found were talks by programmers explaining their methodology and a website that gives an ever so slight background into the matrix models used behind the scenes of a specific rendering engine's algorithm. I have brought these two items into class.

A final note is that Nick, Braden, and myself are in charge of putting together a small presentation on Monday about procedural generation. I will continue to use the material I have found as mentioned above to lay the foundations for this presentations.

XXXIII. July 12th-19th 2013

I just want to write a small log book entry today to show what I have been up to outside of class for the last week. I have been struggling with the analysis of the data that myself and Braden have come up with through our APIE project. Essentially, we have a two dimensional line graph for a given metric. We also have roughly 40 metrics in total and each metric is associated with 11 test projects. This is a lot of data. On top of this we have dates of interest on our line graphs that correspond to project release dates. What we want to ask our data is how the metrics change and evolve over a project's lifetime, specifically around release dates. We have got lots of feedback on how best to proceed with the data analysis, and we have finally, and unfortunately, landed on manual inspection of graph trends. We categorize the graph into 4 types of trends and try to fit in our point of interest into these trends. We believe this will give us a good indication as to how the software evolves around these points. We have roughly set out to do around 2200 manual inspections of our data for this class alone and will continue to move forward into the future.

Edit: I just came back to this post to add another technique we have decided to use. Although we did not use it for the paper for this class, we will be using it for our submitted research paper to MSR 2014. What we did was consider a single test project, a single metric and a single release date. Here, we sum the metrics to the left of the release date by 15, 30, or 60 days then divide by the number of commits. We do the same with the right side of the date. We then create a ratio between the two new normalized metrics. This allows us to see how likely it is for a change to happen before a release or after. We believe these metrics will not only confirm what we saw in our manual inspection but will also provide us with greater insight into a wider span of metrics since the process is a lot faster than manual inspection.

XXXIV. July 22nd 2013

I will just write a quick log book entry for today as most of the material has been covered in previous log book entries. Today, myself, Nick, and Braden presented our material on procedural generation. I have explained in previous log book entries what the material was that I found and how I found it. One interesting thing to note about today was that we all actually found the same material for the most part, we just chose to present different items we found. It is interesting to see how we all had the same mind set when it came to researching the topic which lead to similar results. It is hard to tell if this is a bi-product of the class (learning how to find

research material), the topic being so condensed, or if we are really just the same style of researchers. Just an interesting thought.

A side note to today is that Braden and myself have finished our project for this course and are prepared to present the results on Thursday.

XXXV. July 25th 2013

Today was the first day of presentations in our class about final projects. Braden and myself presented our results today as well as 2 other students. I think our presentation went great with the results actually being of some scholarly caliber. We are looking forward to submitting our results to a real research conference in the near future.

I did not care much for the other two projects presented today as they seemed somewhat easy and fully presented well. It was often difficult to understand what the contribution was to the project.

A final note for this entry is that it seems as though I am now finished my term work for this class as the final project has been presented and the paper has been written. This implies that all future log book entries will simply be about other student's presentations to the class.

XXXVI. July 29th and August 1st

Since all that remained in this class were student presentations, I thought I would aggregate them into one final log book entry and just talk about some of the more interesting ones that I thought were presented.

On the first day of presentations, I really enjoyed Laura's and Nick's presentations. Laura was talking about a tool I had seen published before at a research conference about stack overflow. It was interesting to see it now fully presented with the theoretical background as the main focus. The tool itself is all about the links that are shared on stack overflow and how they inter relate to one another. This obviously leads into a strong graph model back end which is used to find how similar links are together. I think this tool perfectly relates to our class work (especially early on in the class) with graph models and similarity measures. I thought it would actually be interesting to combine some of the results found in Eirini's graph talk with Laura's tool. It would be interesting to see how the two can be combined in order to find the diffusion of links on stack overflow or how they evolve over time.

The second talk I found to be very interesting was Nick's talk on procedurally generated music. Nick came up with his own modified algorithm and actually had fully composed songs by the end. His talk was very clear as to how his algorithm worked and I loved seeing how it actually evolved over time to give better and better music. I feel like the music he generated would be perfect background music for a video game of some sort. Not only does it sound great, but it would also cut down on the size of the game as you would not have to store large mp3 files for music. However, one major issue I saw with this talk (as well as all other talks on the second day) was the lack of comparison.

The issue of comparison was more prevalent on the second day than first but should be discussed for both regardless. When creating a new algorithm or idea, I believe it is of the utmost importance to compare and contrast to previous work. Yes there are cases with grounded theory where this need not apply, but as a scientific community, if we are not comparing our works then we are just spinning our wheels. How will we ever know if method A is better than method B and thus how can we continue to evolve our field of research and become wiser. All 4 presentations on the last day missed out on comparing their ideas to previous works, which in my opinion is a serious flaw. I believe a large benefit to this class and just to our careers would be to have a week or a couple classes dedicated to the scientific process and to how we should be following it not only in our class work but in our own personal research or learning initiatives.