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.