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.

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.