ELEC6211 Project Preparation – Project Plan 1

Title	**Patent Networks: Predictability of Success**
Student name:	**Alun Meredith**
Supervisor name:	**Markus Brede**

Aims/research question and Objectives

Background

The preferential attachment model which has been used to describe patent networks gives the likelihood of a patent receiving a new incoming citation as a function of the number of citations it already has. This 'rich get richer' mechanism describes how already successful patents continue to accumulate new citations. In this model the accumulation of citations in the early growth period is heavily stochastic as there will be many patents of similar age and low numbers of citations, before converging to a more predictable state.

Aim

The aim of this project is to investigate the early growth of a patent and assess the predictability of success within the bibliographic network. Specifically analysing the relationship between predictability and the age of the patent i.e. how old does the patent have to be before the future growth can be reasonably approximated?

Objectives

- **Build a bibliographic network using the US Patent Office Dataset.**
- **Reproduce the work of Valverde *et. al*. estimating the parameters of the preferential attachment model for this network.**
- **Produce a predictive model of the future citation activity of a patent based on the time series of its early growth and implicit parameters such as industry code.**
- **Extract additional features of patents through text mining techniques such as computing measures of semantic distance, building a semantic network and integrating these features into the model.**
- **Evaluate this model, contrasting with preferential attachment. How do the initial features of the patent affect predictability?**

Summary of proposed research and analysis methodology

We propose to build a network model from the dataset through the citations of the patents. This model will assign directed edges between nodes by the citations between patents including encoding the time information when they were formed.

Using this model we may reproduce the work of Valverde *et. al*. focusing on two main aspects.
Firstly the time evolution of number of patents, i.e. how the total number of patents and number of citations per patent has changed over the lifetime of the dataset. This is important because increases in the average number of patents given should be accounted for when training predictions based on historic values.
Secondly the degree distribution, mapped to an extended power law shape and estimation of preferential attachment rule using kernel estimation procedure. Reproducing these two aspects of the work serves as the null hypothesis for our further work. Analysing how the growth of citations deviates from this serves as a strong potential feature in predicting the success of the patent.

At this point additional features can be extracted from the patent data. Technology codes and industry activity are strong potential features to include as patents in industries which see a lot of innovation will be well placed to be cited often themselves, therefore extracting features based on the technology codes and related activity levels of those areas.
Using text mining techniques such as latent semantic analysis different text based features can be extracted, primarily the semantic distance between patent documents which can itself be used as a feature but also used to create a network based on these distances. Patents which live semantically close to other patents could be more likely to succeed because this implies it is situated in an active area but equally those far away could be the root nodes which spawn new areas of research.

We propose to integrate these features into the preferential attachment model of Valverde *et al.* investigating how they affect the underlying parameters of the preferential attachment model. It has been shown that different industries within the patent network can act significantly differently.

Finally we will investigate the growth pattern of citations based on their current time series. This model will compare the growth pattern of patents to the growth pattern described by the preferential attachment model, by comparing a patent to the position it would be in if attachment simply followed the preferential attachment model we can get a sense of the underlying desirability of that patent and predict its success through this implicit desirability.

Research plan – Gantt chart or Pert chart

Ethical statement

There are no pertinent ethical issues regarding the research methods proposed in this project. Regarding privacy, while the dataset being used does contain names of people it is in the public domain and patent applications are by their definition a public declaration.

Regarding the ethical implications of the work presented in the project, there is some potential that the research presented could be used to manipulate the position of patents within the patent network through the potential manipulation of features found as predictors of success. (CopywriteOffice, 2016)

Legal and commercial aspects

Commercial

The potential for commercial use of this research is indirect; using the proposed research to understand the growth of patents within the network can lead to a better understanding of the underlying value of patents and their place within the evolution of innovations. Through this there is potential that innovators can find features of the network that improve their own innovations such as finding forgotten undervalued patents from which to build from or identifying foundation stones of innovation earlier in their lifespan.

Legal

There are very few legal concerns with this research. As the dataset used is open government data, any data protection and privacy issues have been addressed by the curators of the dataset (US patent office, USPTO) while many of these issues aren't pertinent because patents are an open declaration so privacy isn't a concern. There are also no health and safety or environmental responsibility concerns beyond standard computer activity.

The primary legal concern is that the dataset used is owned by the USPTO, however there is an "open licence" for this data allowing the reuse and redistribution of it with no restrictions (CopywriteOffice, 2016)**.**