**P:** It is difficult to predict

**C:** Technology companies, private markets, publicly available data

**S:** A machine learning algorithm

**RQ1:** What are the existing solutions to *predicting whether a company is likely to be acquired*?

**RQ2:** How does the different solutions found by addressing RQ1 compare to each other with respect to analyzing *technology companies*, that are in the *private markets*, using *publicly available data*?

**RQ3:** What is the strength of the evidence in support of the different solutions?

**RQ4:** What implications will these findings have when creating a machine learning algorithm

**Key Terms:**

1. Predicting / Prediction / Regression / Learning / Machine Learning
2. Company / Startup / Venture / Target / Business
3. Acquisition / M&A / Merger / Trade Sale

**1. Problem Definition**

High-growth technology companies (startups) are turning away from the public markets. Amazon went public in 1997, just two years after its first round of institutional financing, at a market capitalisation of \$440M \cite{amazon}. Contrast that with Uber, which remains private six years on and recently raised \$3.5B at a \$59B pre-money valuation \cite{uber}. Time to Initial Public Offering (IPO) for Venture Capital (VC)-backed startups has more than doubled over the past 20 years while VC-backed startups pursuing an IPO has plummeted \cite{nvca2016}. (QUANTIFY)

One explanation for why startups are staying private for longer is the accelerating nature of global business. Startups, particularly those backed by VC firms, are expected to scale fast and require frequent rounds of fundraising coupled with centralized, quick decision making. Such flexibility is not afforded to public companies, due to strict reporting and compliance requirements \cite{wies2015}.

Why does this waiting game matter? Principally, because it shifts value creation to the private markets. To put things in perspective, Microsoft’s market capitalisation grew 500-fold following its IPO \cite{microsoft}, but for Facebook to do the same now its valuation would have to exceed the global equity market \cite{facebook}. VC funding for late-stage startups is approaching all-time highs, possibly because more investors are entering the private markets to seek higher returns \cite{nvca2016}.

Merger and Acquisitions (M\&A) have far surpassed IPOs as the most common liquidity event for startup founders and investors. In 2015, five times as many US-based VC-backed startups were acquired compared to those that went public through an IPO \cite{nvca2016}. Accordingly, startup founders and investors may be interested in predicting which startups are likely to be acquired and by whom. However, M\&A prediction is a challenging task.

Previous work has relied on relatively small data sets \cite{wei2008} because publicly-available information on private companies is scarce. In addition, previous work has focused on the financial or managerial features of potential targets \cite{hongjiu2007} with little work on textual or social network features.

Xiang and colleagues \cite{xiang2012} addressed some of these challenges by mining CrunchBase profiles and TechCrunch news articles to predict the acquisition of private startups. Their corpus was larger than previous studies: 38,617 TechCrunch news articles from June 2005 - December 2011 mentioning 5,075 companies, and a total of 59,631 CrunchBase profiles collected in January 2012. Their approach achieved a True Positive rate of between 60-79.8\% and a False Positive rate of between 0-8.3\%.

There are limitations to Xiang and colleagues' study: the CrunchBase corpus they studied was sparse, only a few common binary classification techniques were tested, and their approach didn't consider IPOs or bankruptcies as potential outcomes. In addition, it is unclear how robust their classifiers are through time. The study could be extended by applying the topic modelling approach to other text corpora such as patent filings, or by attempting a social network link prediction model.

**2. Rationale**

We aim to produce a supervised learning model that will accurately predict the acquisition of startups in the private markets. We will build on the study by Xiang and colleagues (2012) \cite{xiang2012}, introducing new features and classification techniques. In the previous study, True Positive rate (TP), False Positive rate (FP) and Area under the ROC curve (AUC) were the main evaluation metrics used (collectively, known as ``accuracy").

Introducing new classification techniques improves accuracy. Xiang and colleagues' study tested three common binary classification techniques: Bayesian Networks (BN), Support Vector Machines (SVM), and Logistic Regression (LR). BN significantly outperformed SVM and LR. The authors suggested that this was because of the high correlation among their features and absence of a linear separator in the feature space. We will test a number of new classification techniques including Random Forests (RF), CART Decision Trees (CART), and Restricted Bolzmann Machines (RBM), to try to improve the accuracy of the model.

Introducing additional CrunchBase features improves accuracy. Xiang and colleagues' study used a total of 22 factual features from CrunchBase profiles. No feature selection process was documented. A recent similar study on AngelList (which has a sharing agreement with CrunchBase) used 85 features of which 11 were selected \cite{beckwith2016}. Of those 11 features, many were not included in Xiang and colleagues' model. It is plausible that broadening the feature space may result in an improved model.

Introducing additional labels improves accuracy. Xiang and colleagues' study labelled startups as either ``acquired" or ``not acquired". The ``not acquired" category thus includes startups that have bankrupted as well as highly successful startups that went public through an IPO. It is plausible that the breadth of this category would lead to misclassification. Introducing labels for ``public" and ``bankrupt" could improve the accuracy of the model.

Using more recent CrunchBase corpora improves accuracy. Xiang and colleagues' study used a CrunchBase corpus from January 2012. They found the corpus relatively sparse at the time. Since 2012, the CrunchBase corpus has significantly grown. The CrunchBase Venture Program and the AngelList - CrunchBase data sharing agreement have contributed to the corpus, in addition to natural growth over time. It is plausible that a more recent CrunchBase corpus will provide a better basis for a more accurate model.

This study will improve our understanding of the determinants of startup acquisition in the private markets. The system devised by this study also has the potential to de-risk venture capital and encourage greater investment in private startups.

**3. Data Management / Sources**

**4. Feature / Model Selection**

**5. Machine Learning Techniques**