MIT-LL Graph QuBE: Probabilistic Query-by-Example of Large Transactional Data Sets

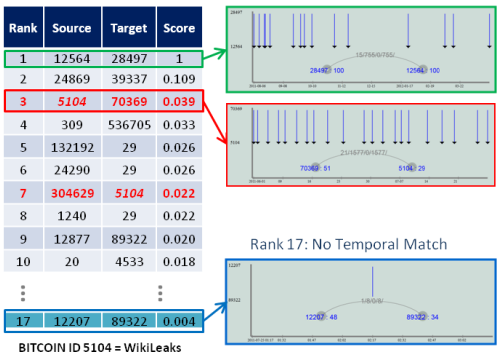
Wade Shen swade@ll.mit.edu
Charlie Dagli dagli@ll.mit.edu
Michael Yee myee@ll.mit.edu
Gordon Vidaver gordon.vidaver@ll.mit.edu
Joseph Campbell jpc@ll.mit.edu

Executive Summary

Knowledge workers specializing in transactional data are often faced with the following information need: "I have an interesting pattern of behavior between some interesting entities. Can I look through all our data to find other such examples of this type of behavior?" Having such a "pattern-of-behavior" search capability could support many organizational goals such as, for example, situational awareness and decision support. Developing technology for pattern-of-behavior search in large, transactional data sets, however, remains a challenging task for both the commercial and academic communities. Patterns-of-interest in this type of data are often localized in time, involve only a few transactions, and are often buried among many other non-interesting behaviors.

Off-the-shelf commercial approaches (i.e. trigger-based alerting mechanisms) often look at behavioral features in aggregate, detecting outlying actors whose behaviors differ as a whole from the rest of the data. These approaches often fail in the needle-in-a-haystack scenario described above. Traditional query-by-example pattern search technologies from academia fail to adequately address the inherent challenges posed by transactional data. Sub-graph search algorithms find structurally similar groupings of actors, but cannot match temporal characteristics. Conversely, powerful temporal matching algorithms are rendered impractical by the exponential number of groupings against which to match.

MIT Lincoln Laboratory is marrying the strengths of these traditional approaches from academia with the practical imperative from industry to develop an efficient two-stage graph query-by-example (Graph QuBE) system for pattern-search on transactional data. Quantitative experiments on the DARPA XDATA 2013 Summer Challenge Data sets show the efficacy and potential of this technology.



Pattern Search: Donation Activity in Bitcoin
Given a pattern of donation activity among suspicious actors in Bitcoin, GraphQuBE pattern-search returns other examples of the same pattern-of-behavior, involving similar entities. The query pattern is visualized in the green box at top. The pattern is a series of donations at regular time intervals. Several patterns returned high on the list (highlighted in red) correspond to patterns involving Wikileaks, a known suspicious actor. Much further down the list (highlighted in blue), we return matches whose temporal pattern is not as similar to the query.

As seen in the figure above, we've shown how recurring donation patterns among suspicious looking entities in Bitcoin can be used to query for similar patterns among similarly suspicious entities, including one returned example involving *Wikileaks*. Additionally, we've demonstrated how the entity-search functionality of Graph QuBE can be used to find high-risk partners from the Kiva micro-loan data set given a known high-risk partner as the query.

Graph QuBE is a freely-available open-source tool open to improvement and/or integration by others. As an example of such an extension, its entity/pattern search capability has been integrated into Influent, a larger "follow-the-money" style visualization tool developed by Oculus, Inc. as part of DARPA's XDATA program.