DARTMOUTH

Institute for Security
Technology and Society
at Dartmouth

Randomized Generation of Adversary-Aware Fake Knowledge Graphs to Combat Intellectual Property Theft

Cristian Molinaro, Andrea Pugliese, V.S. Subrahmanian, Snow Kang

Theft of intellectual property (IP) is a growing problem. Publicly available estimates suggest that over 20% of US companies believe that they have been the victims of intellectual property theft. A 2018 CNN report states that losses due to IP theft by Chinese actors alone cost the US economy somewhere between $225 billion and $600 billion.

Recent Dartmouth-led efforts have proposed combating IP theft by generating fake versions of technical documents. These visionary ideas are based on the policy of deterrence: by ensuring that an enterprise system has *N* fake versions of every real document, the potential victim inflicts costs on the attacker. The IP thief must sift through an array of different versions of a real document, trying to separate the real one from a sea of fakes. This costs the attacker time and money – and inflicts pain and frustration on the part of its technical staff.

***Clique-FakeKG* leverages a graph-theoretic approach to achieve a formalized adversary-awareness standard.**

Recent Dartmouth systems such as FORGE and WE-FORGE achieve these deterrence goals by replacing concepts in an original document with different concepts. However, these past proposals have several limitations: (i) they identify concepts to replace in the document but do not consider the detailed knowledge encoded within a document, and (ii) they do not account for an adversary who both knows the algorithm for generating fakes and can analyze relationships between the resulting set (real + fakes) to find the real one.

In order to tackle these concerns, the Dartmouth College and University of Calabria team created the *Clique-FakeKG* algorithm that leverages a graph-theoretic approach to achieve a formalized adversary-aware standard. That is, our algorithm takes an "original" knowledge graph (KG) K_o that captures some intellectual property and creates fake yet highly believable KGs from K_o, so that after putting them all together, an adversary has no clue which is the original. When testing *Clique-FakeKG* on human subjects, we achieved an 86.8% deception rate with users showing difficulty in distinguishing which KG from a set of KGs was the original K_o from which all the graphs were all generated.

In addition to technical documents, *Clique-FakeKG* can also be applied to arbitrary knowledge graphs such as those used extensively in the finance, retail, and other sectors.

The Fake-KG Problem

We envision our algorithm Clique-FakeKG to be used in a 3-step process to generate fake versions of a technical document: (i) Given an original document, we extract a knowledge graph. (ii) We then use the techniques in our paper to generate a set *K* of fake KGs. (iii) For each of the fake KGs in *K*, we then generate a fake document. *Our work focuses on (ii).*

The algorithm to solve the *Fake-KG* Problem, entails taking an original KG K_o and generating *K*, a set consisting of K_o mixed in with *n* fake KGs. We include another parameter τ, which is an interval that allows users to set a minimum desired distance between every pair of distinct KGs in *K* (meaning fake KGs must be "far enough" from the original one) as well a maximum desired distance between every pair of distinct KGs in *K* (meaning fake KGs must not be "too far" from the original one in order to keep them "believable enough"). The *FakeKG* algorithm also allows users to customize which distance function is used and to provide a set *U* of KGs, which are all the KGs of interest for the application at hand.

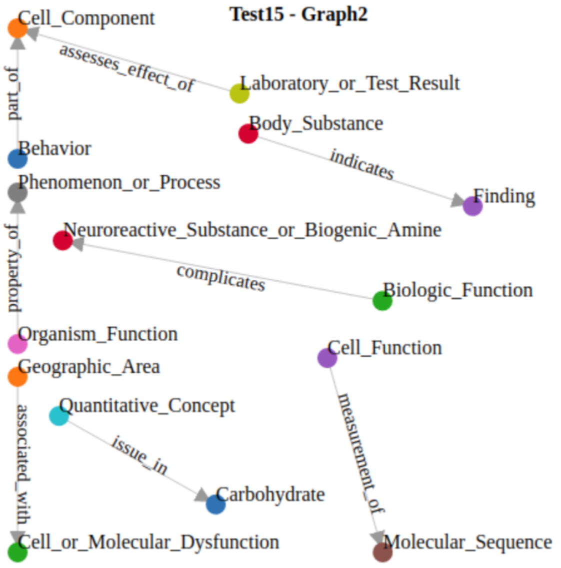
An Adversary-Aware Standard

In developing a strong algorithm to generate *K*, we needed to assume a strong adversary – one that knows exactly how our algorithm works as well as every input the algorithm receives. The only input s/he does not know is which KG out of the set of KGs *K* is the original one (K_o*)*. This "adversary-aware" standard places us in a very hostile setting which requires a carefully designed algorithm. To demonstrate this point, we were able to show that two natural and intuitive algorithms both inadvertently leak information and give the adversary some sense of which KGs are more/less likely to be K_o.

A Graph-Theoretic Approach

Our *Clique-FakeKG* algorithm reexamines the problem from a graph-theoretic lens. If we want to generate *n* number of fakes from an original graph K_o such that the distance between all the graphs (original + fakes) lies within our input interval τ, then we can do the following. First, we create a distance graph between all feasible KGs: vertices in the distance graph are KGs and edges are only drawn between KGs if the distance between them is within our input interval τ. Then, if we find an (*n+1*)-clique in our distance graph that contains K_o, this means we have found *n* fake KGs that fall within the desired distance interval from one another and from K_o.

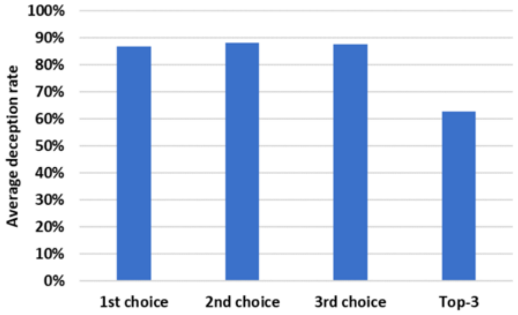
Finally, we showed that *Clique-FakeKG* meets our "adversary-aware" standard by finding maximum cliques to ensure all the KGs in *K* have the same number of solutions. We prove that as a result, each KG in *K* is equally likely to be the original KG and the adversary gains no additional information even after knowing our algorithm, our algorithm's inputs, and the output *K*. Our algorithm also offers several computational benefits, such as storing computed values for subsequent calls of the algorithm and making the computation less demanding as the algorithm proceeds.



A fake graph generated using the Clique-FakeKG algorithm on the UMLS dataset.

High Levels of Deception

We validated the efficacy of *Clique-FakeKG* on 3 diverse real-world datasets: *Nation* which represents international relations among countries, *UMLS* (which represents biomedical relations, and the Microsoft *FB15K-237* (Toutanova et al. 2015), which stores triples and textual mentions of Freebase entity pairs. We built a set *T* of 66 tests in total, each consisting of 10 KGs including the original one. We then asked 10 human subjects (all with a Masters or a Ph.D. in Computer Engineering) to review the 66 tests in *T* and its 660 KGs overall. We used a web-based tool to visualize the KGs as directed graphs with labeled vertices and edges and asked the subjects to select, for each test, the top-3 KGs they felt were the original. We then defined a metric called *Deception Rate (DR)*. For each subject *h*, each original KG K_o, *and r* ∈ {1,2,3,top-3}, we write *w*(*h, Ko, r*) = 1 if a fake KG was selected as the *r*-th choice by *h*, and *w*(*h, Ko,r*) = 0 otherwise—in the top-3 case, we assumed the human subject was correct when any of his/her top-3 choices was right.



Average deception rate for the different choices.

Our experiments show that in 86.8% of the cases ,the KG that users selected as their top choice was in fact fake. Thus, our results demonstrated that **our approach is successful at deceiving users.**

References

S. Kang, C. Molinaro, A. Pugliese, V.S. Subrahmanian. Randomized Generation of Adversary-Aware Fake Knowledge Graphs to Combat Intellectual Property Theft, *Proc. AAAI – 2021,* Feb 2021.

Data & Code

https://github.com/snow-kang/FakeKG

Please acknowledge our paper if you use our data/code.

PARTICIPANTS

Lead: V.S. Subrahmanian

Cristian Molinaro, Andrea Pugliese, Snow Kang

Dartmouth
Security and Artificial
Intelligence Laboratory