Characterizing Mosaicing Inference Risk: A Preliminary Study

Nathaniel Rollings
University of Maryland
College Park, MD, USA
nrolling@umd.edu

Douglas W. Oard
University of Maryland
College Park, MD, USA
oard@umd.edu

Abstract

As information proliferates, inference risks for as-yet unstated facts that require protection rise. This has been called the "mosaicing" challenge. The costs of manual mosaicing review limit what can be reviewed, leaving substantial quantities of actually innocuous information unreviewed, unreleased, and thus unsearchable. This paper models mosaicing risks in two ways: (1) as multi-hop question answering in text and (2) as relation inference in knowledge graphs. However, our focus is not on inference, but rather on inference prevention. In each case, this is done by dividing the information space into two parts, a larger part already widely known, and a smaller review set being considered for potential public disclosure. The goal for a system is to identify cases in which facts that require continued protection would be at increased risk of inference if some item in the review set were disclosed. Results of experiments show some protection can be achieved with currently known techniques, but that inference ability may not predict inference prevention well.

CCS Concepts

• **Information systems → Information retrieval**.

Keywords

Information protection, Question answering, Knowledge graphs

ACM Reference Format:
Nathaniel Rollings and Douglas W. Oard. 2025. Characterizing Mosaicing Inference Risk: A Preliminary Study. In *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR) (ICTIR '25), July 18, 2025, Padua, Italy.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3731120.3744577

1 Introduction

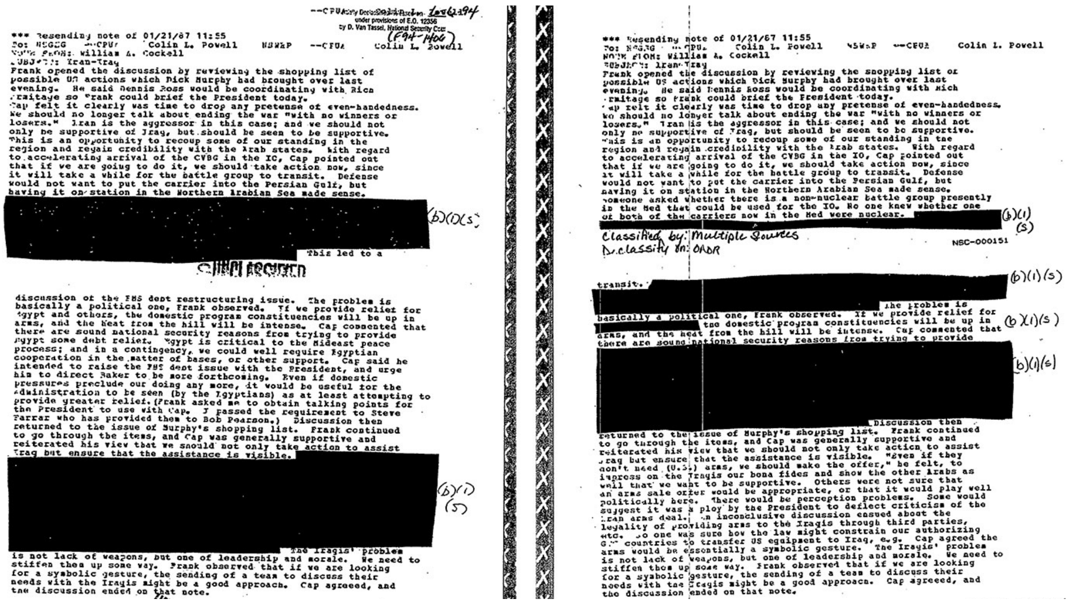
Our ability to search relies on content having first been shared, and thus made searchable. There is, however, much that people are reluctant to share for many reasons, including commercial advantage, personal privacy, or national security [3, 33, 49]. Sensitive content is often intermixed with innocuous content that could, if separable, be freely shared. That has led to a line of work in information retrieval on detection and removal of sensitive content before indexing [10, 28], and on search engines that can segregate sensitive content at query time [40]. Almost all such work has relied on a document independence assumption, making decisions on what is sensitive for each document in isolation; the one exception we know of groups related documents to allow them to be considered together [31]. In this paper, we push further on developing automated tools to help reviewers identify inference risks.

Our work is motivated by what has been called the "mosaicing" problem in declassification, in which the goal is to determine when content being considered for public release would enable inference of secrets that require continued protection [37, 38]. Finding cases where individually releasable facts could together allow undesirable inferences can be challenging. A same-document mosaicing case is illustrated in Figure 1. Here we see two sets of redactions to a classified message sent from William Cockell to Colin Powell in 1987 [32]. While some text is redacted in both versions, different redaction decisions permit some inference about the secret(s) being protected. For instance, the left version of the document includes information about how a previous statement (redacted in that version) "led to a discussion of the FMS debt restructuring issue" involving Egypt. On the right, by contrast, references to Egypt are redacted, but not the question about "whether one or both of the [aircraft] carriers now in the Med[iterranean Sea] were nuclear." When we put this together with publicly known information, the reason for the redactions becomes clearer. A couple of years earlier, the Washington Post had reported that "Egypt allowed a nuclear-powered U.S. Navy ship to pass through the Suez Canal for the first time last weekend in what one official called a 'breakthrough' for U.S. diplomacy" [20]. What we see in this message is a discussion of how a U.S. offer of debt relief to Egypt might help to secure permission for nuclear-powered ships to pass through the Suez Canal.

It is straightforward to prevent same-document mosaicing attacks by simply finding prior redaction decisions for the same document and calling them to the attention of a human reviewer who is making future redaction decisions [41]. Cross-document mosaicing attacks pose much greater challenges because we must search all extant information for tidbits that could be used, in combination with information in a document being considered for release, to infer a secret. Mosaicing is further complicated by the inability to retract information that has already been released, even if that substantially reduces the challenge of inferring sensitive information [1, 37]. We ultimately seek to develop a framework that can identify and reduce mosaicing risks that have the potential to help people perform mosaicing review tasks. Our first step toward that goal is to identify when one system for nominating potential redactions is better than another. That is our focus in this paper.

We investigate the use of current techniques to characterize inference risk in both text and knowledge graphs. We divide information into two sets on which our inference models operate: (1) a public set that represents what is already publicly known and thus must be accommodated and (2) a review set that represents information not yet public. The attacks we seek to prevent are those

ICTIR '25, July 18, 2025, Padua, Italy
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1861-8/2025/07
https://doi.org/10.1145/3731120.3744577



--CFUA Partially Declassified/Released on 10/6/94 under provisions of E.O. 12356 by D. Van Tassel, National Security Council (F94-1406)

*** Resending note of 01/21/87 11:55
To: NSGCG --CPUA Colin L. Powell NSWRP --CFUA Colin L. Powell
NOTE FROM: William A. Cockell
SUBJECT: Iran-Iraq

Frank opened the discussion by reviewing the shopping list of possible US actions which Dick Murphy had brought over last evening. He said Dennis Ross would be coordinating with Rich Armitage so Frank could brief the President today.

Cap felt it clearly was time to drop any pretense of even-handedness. We should no longer talk about ending the war "with no winners or losers." Iran is the aggressor in this case; and we should not only be supportive of Iraq, but should be seen to be supportive. This is an opportunity to recoup some of our standing in the region and regain credibility with the Arab states. With regard to accelerating arrival of the CVBG in the IO, Cap pointed out that if we are going to do it, we should take action now, since it will take a while for the battle group to transit. Defense would not want to put the carrier into the Persian Gulf, but having it on station in the Northern Arabian Sea made sense. [redacted] (b)(1)(s) This led to a

UNCLASSIFIED

discussion of the FMS debt restructuring issue. The problem is basically a political one, Frank observed. If we provide relief for Egypt and others, the domestic program constituencies will be up in arms, and the heat from the hill will be intense. Cap commented that there are sound national security reasons from trying to provide Egypt some debt relief. Egypt is critical to the Mideast peace process; and in a contingency, we could well require Egyptian cooperation in the matter of bases, or other support. Cap said he intended to raise the FMS debt issue with the President, and urge him to direct Baker to be more forthcoming. Even if domestic pressures preclude our doing any more, it would be useful for the administration to be seen (by the Egyptians) as at least attempting to provide greater relief. (Frank asked me to obtain talking points for the President to use with Cap. I passed the requirement to Steve Farrar who has provided them to Bob Pearson.) Discussion then returned to the issue of Murphy's shopping list. Frank continued to go through the items, and Cap was generally supportive and reiterated his view that we should not only take action to assist Iraq but ensure that the assistance is visible. [redacted] (b)(1) (s) The Iraqis' problem is not lack of weapons, but one of leadership and morale. We need to stiffen them up some way. Frank observed that if we are looking for a symbolic gesture, the sending of a team to discuss their needs with the Iraqis might be a good approach. Cap agreeed, and the discussion ended on that note.

*** Resending note of 01/21/87 11:55
To: NSGRG --CPUA Colin L. Powell NSWRP --CFUA Colin L. Powell
NOTE FROM: William A. Cockell
SUBJECT: Iran-Iraq

Frank opened the discussion by reviewing the shopping list of possible US actions which Dick Murphy had brought over last evening. He said Dennis Ross would be coordinating with Rich Armitage so Frank could brief the President today.

Cap felt it clearly was time to drop any pretense of even-handedness. We should no longer talk about ending the war "with no winners or losers." Iran is the aggressor in this case; and we should not only be supportive of Iraq, but should be seen to be supportive. This is an opportunity to recoup some of our standing in the region and regain credibility with the Arab states. With regard to accelerating arrival of the CVBG in the IO, Cap pointed out that if we are going to do it, we should take action now, since it will take a while for the battle group to transit. Defense would not want to put the carrier into the Persian Gulf, but having it on station in the Northern Arabian Sea made sense. Someone asked whether there is a non-nuclear battle group presently in the Med that could be used for the IO. No one knew whether one or both of the carriers now in the Med were nuclear. [redacted] (b)(1) (s)

Classified by: Multiple Sources
Declassify on: OADR

NSC-000151

transit. [redacted] (b)(1)(s) The problem is basically a political one, Frank observed. If we provide relief for [redacted] (b)(1)(s) the domestic program constituencies will be up in arms, and the heat from the hill will be intense. Cap commented that there are sound national security reasons from trying to provide [redacted] (b)(1)(s) Discussion then returned to the issue of Murphy's shopping list. Frank continued to go through the items, and Cap was generally supportive and reiterated his view that we should not only take action to assist Iraq but ensure that the assistance is visible. "Even if they don't need (U.S.) arms, we should make the offer," he felt, to impress on the Iraqis our bona fides and show the other Arabs as well that we want to be supportive. Others were not sure that an arms sale offer would be appropriate, or that it would play well politically here. There would be perception problems. Some would suggest it was a ploy by the President to deflect criticism of the Iran arms deal. An inconclusive discussion ensued about the legality of providing arms to the Iraqis through third parties, etc. No one was sure how the law might constrain our authorizing GCC countries to transfer US equipment to Iraq, e.g. Cap agreed the arms would be essentially a symbolic gesture. The Iraqis' problem is not lack of weapons, but one of leadership and morale. We need to stiffen them up some way. Frank observed that if we are looking for a symbolic gesture, the sending of a team to discuss their needs with the Iraqis might be a good approach. Cap agreeed, and the discussion ended on that note.

Figure 1: A simple mosaicing example [32]. The same document was released twice, with different redactions.

made by inference models against our secrets. We call these models evaluators when used for inference. When used for redaction (information removal), we refer to a model as a "nominator." Our principal contributions are formalizing the mosaicing inference task, presenting comparable evaluation measures for that task in text and knowledge graphs, and illustrating the impact of redactions on inference models.

2 Related Work

Sensitive Information Protection. Inference prevention has been a focus in privacy protection, so we begin there. Considerable work exists on strategies to limit unique identification of individual entities in a database [2, 26, 43]. Some, notably k-anonymity and l-diversity, have been applied to social media graph anonymization [9]. Differential privacy also protects sensitive information [11, 21], but its addition of noise is inconsistent with our focus on redaction rather than obfuscation. We have also seen work on detecting sensitive documents in IR, as noted above. In graphs, we see substantial work on manipulating data to protect privacy. Casas-Roma et al. investigated edge manipulation in unlabeled and undirected social media graphs [8]. Qian et al. leverage node similarity for inference in unlabeled graphs, a specialized case of graph completion [39]. Efforts at privacy protection within knowledge graphs exist any inference of a specific type of relation [13], while our efforts focus on preventing specific inferences.

Question Answering. We are particularly interested in Multi-Hop QA (MHQA), in which the answer to a provided question is not directly available in the collection but instead must be inferred from information that may be found in multiple documents [27]. This can include inferential reasoning across a large collection, often with retrieval and inference components [50, 51]. We focus on inference in this paper. For inference in MHQA, we can usefully divide approaches into LLM and non-LLM methods. Both generative LLMs [23, 25] and large BERT-derived language models [16, 17] see extensive use. Implementations vary, though some approaches generate small graphs over which they then reason [16, 17, 25], offering some potential for bridging our work on text and knowledge graphs in the future. Generative approaches are particularly interesting given the upsurge in investment in extremely large models of this type. While LLM-based approaches have attracted attention in recent years, the present state of the art on the HotPotQA dataset with which we experiment is Beam Retriever, a non-LLM model [51]. Another non-LLM model we looked at used a series of bi-directional GRUs to process text in the context of a question [14]. That model was used (with a few modifications) as a baseline for the HotpotQA dataset experiments [50]. We have used both of these models in our experiments, and describe them in more detail below.

Knowledge Graph Completion. Knowledge graphs are networks of entities with directed edges denoting relationships between pairs of entities, typically encoded as (subject, relation, object) triples [39]. The completeness of these graphs is limited, however, so the Knowledge Graph Completion (KGC) task attempts to identify missing relations in such graphs [12]. KGC can be divided into two broad categories: transductive and inductive. Transductive approaches are trained and tested on the same graph, often by learning embeddings for triples in the graph [42, 47, 48]. Inductive models, by contrast, require only that the same types of relations be present in the training and test graphs. KGC models range from simple statistical approaches [29] to more complex models designed to identify distinctive aspects of the neighborhood of relations surrounding a target relation [44]. We focus our work on inductive models because these models often outperform transductive models, even on transductive tasks [44, 52], and because transductive models could require retraining after any proposed redaction.

Adversarial Knowledge Graph Completion. The closest work to ours in knowledge graphs of which we are aware is that of Pezeshkpour et al. [35] on adversarial knowledge graph completion. In that work, they experiment with a restricted class of highly local redactions

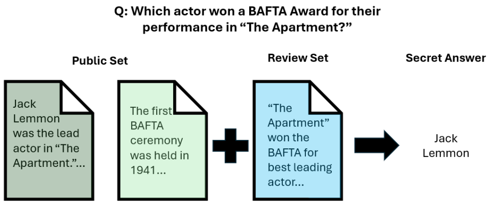

Q: Which actor won a BAFTA Award for their performance in "The Apartment?"
Public Set
Review Set
Secret Answer
Jack Lemmon was the lead actor in "The Apartment."...
The first BAFTA ceremony was held in 1941...
"The Apartment" won the BAFTA for best leading actor...
Jack Lemmon

Figure 2: We seek to protect Jack Lemmon having won Best Actor, but his being lead actor in "The Apartment", and "The Apartment" winning best lead actor, causes inference risk.

in which the triples to be protected, and the triples that can be redacted, share a common object entity. Our setting is more general, allowing redaction of any relation along a path between a secret triple's subject and object entities. In practice, only 19% of the triples we redact in our experiments in Section 4 share a common object with the secret we are seeking to protect. Although we focus on relation removal, other adversarial KGC strategies have also been proposed. Some, such as relation addition [5], entity addition [7], and relation modification [8] would be inconsistent with our interest in adding only truthful knowledge graph content while protecting our secret. Others, such as entity aggregation [19] and other forms of clustering [22], could be consistent with that goal, but there is limited scope for application of such techniques in our task because we treat the large portion of the knowledge graph that is already public as immutable. With these efforts on adversarial graph modification [35], we can identify some overlaps in goals and terminology. However, we also introduce some elements of our own terminology to better describe our specific task. The "target" based terminology of adversarial KGC fails to highlight our reason for preventing inference of this target: its sensitivity to the information provider. In a similar manner, withholding information does not fit precisely with the "attack" and "poisoning" terminology prevalent in work on adversarial KGC.

3 Sentence Redaction in Text

To simulate the problem in text, we started with a well-known multi-hop QA dataset: HotpotQA [50]. We randomly selected 4,000 question-answer pairs to represent secrets, and we arbitrarily selected one of the two "gold" paragraphs in the set of ten dataset-provided paragraphs as the review set.[1] We used the other nine paragraphs to represent potentially relevant content from the public set that we cannot redact. We then employed some inference model to attempt inference of the secret answer over the full set of ten paragraphs (i.e., the union of the one-paragraph review set and the nine paragraphs from the public set). Figure 2 illustrates this setup. For each secret answer, we exhaustively applied single-sentence ablation to identify which sentences in the review set (if any), when redacted singly, would prevent correct inference of the secret answer by that same model. We call such a sentence a model's *nomination* for redaction. We then deleted a single sentence nominated by one model and measured the effect of that redaction on inference of the secret answer by another model. We call that first model the *nominator* and the second the *evaluator*.[2]

3.1 Methodology

HotpotQA was built using Wikipedia abstracts by providing a starting abstract and the abstracts of pages linked from the starting abstract [50]. It is structured like many other multi-hop QA datasets; for inference, the dataset presents a system with a set of Wikipedia abstracts (ten in this case), at least two of which contain information needed to answer an associated question. HotpotQA has 90,447 training and 7,405 validation examples, each with a question, the correct answer, and ten paragraphs, two of which are marked as the "gold" paragraphs that were used to produce the answer during dataset creation. The dataset's creators do not provide a test set and instead take submissions of systems for an unseen test set that is not available to the public. We therefore sample our 4,000 secrets from the validation set.

We employ four inference models for our experiments; two LLMs (GPT-4o [34][3] and Llama 3.0 70B [46]) and two non-LLM models (Beam Retriever [51] and the baseline RNN model from the initial HotpotQA paper [50]). Each LLM is provided one of three prompts, each of which asks the model to answer the provided question using only the given ten paragraphs and to explain its reasoning. The first prompt (1P) designates the review set and asks for a single sentence from that paragraph that helps explain the model's response. The second (2P) is nearly identical but asks for two such sentences, both from the review set. The third (2A) is similar to 2P, but does not designate the review set. We tried a number of other prompts, ranging from slight variations in wording to working entirely with JSON, but these three prompts provided the best results, both in correct answers and in identifying sentences for redaction.

Our first non-LLM model is Beam Retriever [51] ("beam" for short), which is the current state-of-the-art for HotpotQA. It encodes and scores different subsets of the context, using DeBERTa and looking at the top k beams to identify the most useful paragraphs, and then employs DeBERTa again for encoding and start/end answer token detection. Our final model, "rnn," consists of a series of interconnected RNNs; this was a baseline in the original HotpotQA paper [50]. Our models span a range of techniques, with different levels of exposure during training to the Wikipedia content from which HotpotQA was built (rnn: none, beam: only from DeBERTa training, LLMs: extensive). In addition to these existing models, we implemented a simple answer-matching baseline nominator, which searches for the answer in the review paragraph and lists each of those sentences as possible nominations. Note that it is incapable of serving as an evaluator since it lacks inferential capabilities.

For our experiments, we applied the models as nominators and evaluators in every possible combination. We redacted each sentence in the review set in turn, and noted which changed the nominator's answer from correct to incorrect. There are typically about four sentences in the paragraph that we use to model the review set, so this is not too expensive for our experiments (although with large review sets it would be costly). We call this process exhaustive search for sentence redaction. If we find more than one redaction

[1] Gold paragraphs are known to suffice for inference; other sets may also suffice.

[2] All code and data for this paper is available at https://github.com/nroll38/mosaic_ictir

[3] We used the 2024-08-06 GPT 4o checkpoint.

that changes the answer from right to wrong, we arbitrarily select the first to appear in the review paragraph as the nominated redaction. While our approach simplifies the task to single-sentence redaction rather than including cases where multi-sentence redactions are necessary, it provides insight into the viability of our approach to this problem while avoiding the factorial costs in the multi-redaction setting.

When we remove a nominated sentence from the review set, we replace it with "[REDACTED]," in a manner reminiscent of government redaction of classified information. We then run each evaluator model on both the unredacted and redacted versions of the context, and we run our LLMs with every prompt. A nominated redaction is successful for a given evaluator if, when the nominated sentence is present, the evaluator is able to correctly answer the question, but when it is redacted, the model can't do so.

There are two special cases that we need to consider. Sometimes a model simply cannot answer a question correctly, whether redaction is performed or not. For the state-of-the-art beam model, that happens for about one-third of all questions (specifically, 1,278 of our 4,000-question sample). We cannot use such questions to evaluate the effect of redaction. The other special case is when the model persists in answering the question no matter which single sentence we redact from the review set. This can happen for at least three reasons: (1) the model answers from prior knowledge rather than from the paragraphs we provide, (2) more than one sentence in the paragraph that we choose as the review set suffices for the inference, or (3) no sentence in the review set is needed to perform the inference because information provided in the other paragraphs suffices. We can identify these special cases by simply using the same model as both nominator and evaluator. When a question can be answered by a model before redaction, we say that the question has "Exposure" to that model. When a single-sentence redaction can flip an answer to such a question from right to wrong, we say that the question exhibits "Reducible Exposure," since redaction can reduce the Exposure. Note that both Exposure and Reducible Exposure can differ for different evaluator models.

3.2 Results

We first address model performance as evaluators, and then investigate their capability as nominators.

3.2.1 Evaluators. A strong evaluator model would be one that can answer many questions correctly, and whose ability to answer questions is robust to redaction of individual sentences from the paragraph that forms the review set. In other words, for a strong evaluator, Exposure (called "Total Exposure" in our results tables) should be high and Reducible Exposure should be low. As Table 1 shows, beam is better than any other model at answering questions (producing the exact answer that the HotpotQA collection specifies in 2,722 of 4,000 cases), but rnn is far more robust to single-sentence redaction (with the lowest Reducible Exposure, just 261 of the 4,000 questions). As a single figure of merit, we can characterize the inference risk from an evaluator as the difference between Exposure and Reducible Exposure, which is the portion of the 4,000 cases in which that evaluator model can produce an answer that no single sentence redaction from the review set could prevent. By that measure, beam is clearly our strongest evaluator, making "Irreducible" inferences in 2,083 of 4,000 cases.

We see that every GPT prompt outperforms any Llama prompt, with both higher Exposure and lower Reducible Exposure. We also experimented with LLaMA 8B (not shown), finding even lower Exposure and Irreducible Exposure than Llama 70B. While these few experiments are not conclusive, they do support our expectation that larger models are better able to conduct inference, and that inference by larger models is more robust to changes in the set over which the inference is performed. With the LLM prompts, there is a clear tendency toward higher Exposure and lower Reducible Exposure when the review set is specified in the prompt. This is true for both GPT and Llama, regardless of whether we ask for one or for two sentences that support the model's stated inference.

For the results in Table 1, we used exhaustive search to select the sentence in the review set to redact, but in a large-scale application, we would need a more targeted approach. However, we find that when we prompt an LLM to suggest which sentence in a specific paragraph to redact, it is no better than random. Specifically, G1P (GPT with the 1P prompt) suggested a protective redaction for 54% of the cases in which protection was possible (530 of 990), and L1P also made a good suggestion in 54% of such cases. When we ask the model to suggest two sentences, G2P made a good choice 49% of the time and L2P did so 62% of the time, slightly edging out random guessing. G2A did only slightly worse than G2P at finding a good sentence to redact in the review set (43% vs. 49%), despite having to consider ten times as many sentences.[4] The beam model can also identify supporting sentences, and it recommends a good redaction in the review set in 94% of possible cases (602 of 639). Together, these results indicate that while asking LLMs to nominate sentences might scale reasonably well to larger review sets, systems designed specifically for this inference task are currently much better at suggesting specific sentences.

3.2.2 Cross-Model Nomination. So far, our analysis has focused on finding strong evaluators when the nominator and evaluator are the same model (and for LLMs, the same prompt). In reality, however, we want to find nominators that can reduce the inference abilities of many evaluators, because we can't be sure which evaluator model might be tried by someone who wishes to perform inference. This cross-model nomination scenario is the focus of the right side of Table 1. There, we use percentages to show the fraction of the Reducible Exposure for each evaluator that can be achieved by any particular nominator. That fraction is always 100% when the nominator and evaluator are the same, so we omit those 100% values on the main diagonal for improved readability. Notably, we immediately see that our answer-matching baseline is never the optimal nominator, despite having access to information (the answer) that no other model is provided.

We see that other models struggle to achieve large reductions in Exposure for beam. To take a specific example, when beam is the evaluator, G1P reduces its Exposure less than half as well as beam itself could have done (specifically, G1P reduces the Exposure of beam by 285, which is 44.6% of 639), and no other nominator does even that well. Just examining the bolded values (the highest

[4] For L2A vs. L2P the same comparison is 58% vs. 62%.

Table 1: Model effectiveness for 4,000 exact match answers with exhaustive search for sentence redaction. Total Exposure is fraction of secrets inferred without redaction; Reducible Exposure is maximum possible absolute reduction in Exposure (achieved using the evaluator model as nominator). Percentages show fraction of Reducibility achieved by some other nominator. G: GPT-4o, L: Llama 3 70B, 1 or 2: number of requested sentences, P (present) or A (absent): whether review paragraph is specified in prompt. A * indicates a significant ($p > 0.05$) improvement by the binomial test over all prompts using the same model while a † indicates significance over all other models, but not necessarily over different prompts of the same model.

				Nominator's Fraction of Evaluator's Reducible Exposure↑								
	Total	**Irreducible**	**Reducible**	**Answer**	GPT			Llama			Non-LLM	
Evaluator	**Exposure↑**	**Exposure↑**	**Exposure↓**	**Match↑**	**G1P**	**G2P**	**G2A**	**L1P**	**L2P**	**L2A**	**beam**	**rnn**
G1P	2370	1380	990	34.8%	–	**60.1%***†	50.8%	41.9%	40.1%	36.8%	28.9%	11.0%
G2P	2330	1454	876	37.7%	**68.5%***†	–	59.8%	43.8%	41.9%	37.7%	32.3%	12.2%
G2A	2255	1415	840	38.2%	61.8%	**64.3%**†	–	43.6%	41.3%	37.6%	32.4%	12.0%
L1P	1935	903	1032	38.3%	44.1%	41.2%	38.8%	–	**57.8%***†	50.8%	24.6%	10.4%
L2P	1845	808	1037	36.9%	41.4%	38.2%	36.6%	56.8%	–	**71.2%***†	24.5%	10.2%
L2A	1748	781	967	38.0%	42.1%	38.8%	37.6%	55.5%	**78.3%***†	–	24.7%	10.0%
beam	**2722**	**2083**	639	34.0%	**44.6%**†	42.6%	40.7%	34.0%	35.4%	32.2%	–	15.2%
rnn	1697	1436	**261**	38.3%	**42.5%**	42.1%	39.5%	37.2%	38.4%	34.5%	38.7%	–

percentage achieved by any nominator against that row's evaluator) we see a block diagonal structure for the LLMs: when any GPT prompt is the evaluator, some other GPT prompt is the best other nominator, when any Llama prompt is the evaluator, some other Llama prompt is the best other nominator. In other words, the difference between LLMs is greater for this purpose than is the difference between these prompts. We also see that beam and rnn are poor nominators across the board, a combination of differing (beam) and anemic (rnn) redactions. Because LLMs and beam choose different sentences to redact, we might, in future work, also want to look at ensembling LLM and non-LLM models in an effort to improve the Reducible Exposure for an unknown evaluator.

Diving deeper into the LLM results, we see that G1P is the best GPT prompt as a nominator, except when the evaluator is G2A (where G2P does non-significantly better). We see something similar with Llama, where L1P does best as a cross-model nominator in most cases, but it differs in being significantly worse at nominating against other Llama models. Nonetheless, the consistency of this result leads us to speculate that present LLMs do better with shorter inference chains, so asking for one sentence in support may be a bit better than asking for two for the cross-model case.

3.2.3 Overall Remarks. Our experiments with HotpotQA provide several useful insights into the characterization of mosaicing inference risk. First, measuring Reducible Exposure is a useful way of characterizing that risk. When used in a nominator/evaluator framework, we can identify the strongest inference model, upper bound the effectiveness of redaction against a particular evaluator, and measure the relative effectiveness of other redaction nomination models. We also observed differences between specific models, and while improved models in the future can change specific results, we expect that this evaluation framework will remain useful.

While our experiments with text offer useful insights, there are several reasons why we might prefer a more highly structured setting for further experiments. First, although HotpotQA matches our needs to a useful degree, its design requires that we encode a secret as a question-answer pair. There may be many ways of creating such pairs, but HotpotQA's structure led us to model each distinct secret using just one question-answer pair. Second, in HotpotQA each question-answer pair is independent. But in reality, secrets may cluster; Maxwell Smart had many more secrets than did Charlie Brown. Modeling that kind of clustering might be useful because protecting secrets might be easier (or harder!) if they tend to be near each other in some sense. Third, though text bodes well for realism (since many things that need to be reviewed for declassification are written text), it poses a number of confounds, such as repetition, redundancy, and synonymy, that conspire to complexify our models despite our desire for tighter experimental controls. Finally, inference on text brings its own challenges, including the training set leakage we have highlighted in LLMs and, to a lesser degree, in beam. For these reasons, we find it useful to augment what we have learned in text by exploring another approach to mosaicing in a more controlled setting, knowledge graphs.

4 Relation Redaction in Knowledge Graphs

We model the redaction problem in knowledge graphs as follows: given an existing knowledge graph representing public information, a set of new triples being considered for release, and a secret relation that is in neither of those sets (depicted in Figure 3), our task is to redact the one relation in the review set that provides the most support for inference of the secret. We call this the Most Dangerous Relation (MDR). This approach allows us to identify the most useful potential single redaction. Note the difference here from the setup in text: we ask not whether a redaction prevents a correct inference, but rather whether it makes that correct inference less likely.

4.1 Methodology

To simulate the problem, we start with a dataset widely used in KGC research: FB15k-237 [45]. We first select 150 subject, relation, object triples from among its training set relations to be our secrets and then randomly sample one-third of the remaining training set relations as our review set from which we may elect to redact a relation to protect a secret; the remaining two-thirds of the dataset's training set is the immutable public knowledge graph. We then train three KGC models on only the public graph and ask each trained model to perform inference over the union of the public graph and the review set in an attempt to infer a secret. For each secret, we use ablation to determine the one relation in the review set on which each KGC model most depends when making this inference (i.e., the MDR), and we consider that MDR to be that model's nomination for redaction. We then delete the relation nominated by that one model, and measure the effect of that deletion on inference of the secret by each evaluator model.

4.1.1 Dataset. FB15k-237 was built from FB15k by removing inverse links. For instance, if *A* has a *married to* link to *B*, there may be an inverse link indicating *B* is married to *A*. While inverse links are natural, removing them requires our models to make more nuanced inferences. We use the training set of the FB15k-237 V1 inductive split, which has 4,245 triples and 180 relation types [44]. As FB15k-237 is commonly used in KGC research, it provides a base on which the performance of all our models has been characterized.

4.1.2 Evaluation. Each of our models can score any triple formed by adding a new relation between two existing entities, and we can use these scores to rank any set of triples. We form such a set by adding our secret triple to 100 other confounds found using negative sampling, (presumably incorrect) triples that are also not present in the graph. If an inference algorithm ranks our secret above all of the confounds, we consider it fully exposed. If some of the confounds rank above it, that's good. The more the better. As a measure of Exposure, we use the reciprocal rank of the secret in the list of 101 triples. Negative sampling is commonly used for KGC evaluation [36, 44, 52], stemming from early work on Noise-Contrastive Estimation [18, 30]; it was called "uniform negative sampling" in the seminal TransE work on knowledge graph embedding [6].

We perform negative sampling by randomly selecting 50 triples, each with the actual secret's relation and object, but a random incorrect subject, and another 50 with the secret's actual subject and relation, but a different object. Random selection results in many improbable triples among the set of confounds, but we see that as a feature, not a bug, since ranking an improbable confound ahead of our secret indicates strong protection. For example, convincing a model that a book is more likely to have won the Best Actor award than is the actor Jack Lemmon would indicate Jack's secret is safe.

Our goal is not just to measure how well a KGC model performs inference, but, importantly, how well we can use some KGC nominator model to prevent inference of a secret by some KGC evaluator model. Just as in text, we measure the nominator's effectiveness as Reducible Exposure against a particular evaluator. The Reducible Exposure is the reduction in the reciprocal rank of our secret (since lower reciprocal rank indicates a better protected secret), as measured by the evaluator model, after removing the MDR identified by our nominator model. Given a relation nominated for removal, we run the evaluation model against the review set and the public graph, but with the nominated link removed from the review set, much as we did in our text experiment. Here, however, we observe the reduction in reciprocal rank of the secret in the evaluators' rankings of the secret among the confounds rather than just observing whether or not the observer answers a question correctly, as we did in text. Specifically, we compute Exposure as the mean reciprocal rank of the secret among the confounds for the evaluator model before redaction, and the Reducible Exposure as the mean reduction in reciprocal rank that results from redaction of the MDR:

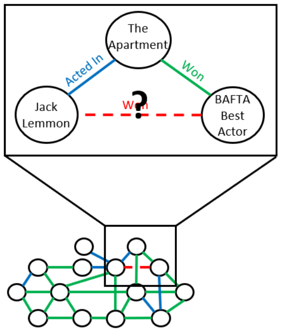


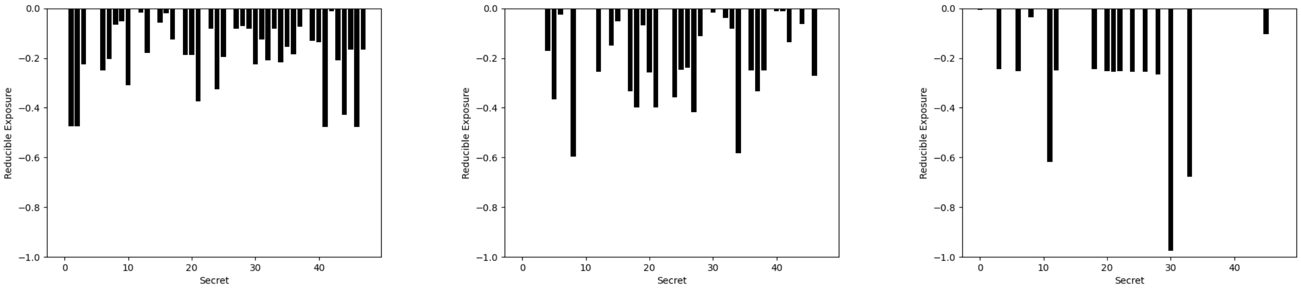
Figure 3: Some relations in a knowledge graph. Red are secret, green are public, and blue are review set. We seek to protect against the same inference as in Figure 2.

$$Exposure = \frac{1}{k}\sum_{i=0}^{k}\frac{1}{r_{e_i}} \qquad ReducibleExposure = \frac{1}{k}\sum_{i=0}^{k}\frac{1}{r_{e_i}} - \frac{1}{r_{n_i}}$$

where r_{e_i} is the rank of secret i for evaluator e without redaction, r_{n_i} is the rank of secret i for evaluator e after redaction of the MDR from nominator n, and k is the number of secrets (150). As with text, Reducible Exposure is maximized when the nominator and evaluator are the same. To see how well other models do, we compute the fraction of same-model Reducible Exposure that some other nominator can achieve, again expressing that as a percentage.

As an example, consider the graph in Figure 3. If we seek to measure the degree to which we can protect the (Jack Lemmon, Won, BAFTA Best Actor) secret, one confound might be (Jack Lemmon, Won, Cleveland). If our secret triple was initially at rank one, but after the nominated single-relation redaction from the review set it drops to rank two, behind this one confound but ahead of all other confounds, the reduction in reciprocal rank would be 0.5. If, however, the secret was initially at rank 10, with nine confounds ranked above it, a further reduction of one rank to rank 11 would produce a reduction in reciprocal rank of just 0.009.

4.1.3 Secret Selection. We start by randomly selecting relations as potential secrets. Of course, random selection may select some "secrets" that would provide little insight into our approaches to inference prevention, in part because random selection may not model the properties of actual secrets well (see section Section 4.3 for experiments with an alternative approach to secret selection). A potential secret that all models can infer using only the public graph will result in Irreducible Exposure since we cannot withhold information that is already public. This happened in 21 of the 150 cases. Second, if no model can infer the secret when using both the public graph and the review set together, before any redaction, then the secret is already safe, at least against the models we have tested, and no further improvement is possible. This happened in 46 of the 150 cases. Finally, if no single relation redaction from the review set can prevent inference of a secret by any model, then multiple redactions would be required, and that is outside the scope of our experiments in this paper. This happened in 35 of the 150 cases. Of the 150 randomly selected potential secrets, 48 passed all three



(a) RuleN (39 of 48 secrets)

(b) GraIL (31 of 48 secrets)

(c) Bellman-Ford (17 of 48 secrets)

Figure 4: Reducible Exposure per secret for each evaluator. The 48 secrets are in increasing order of relation type frequency.

filters. We used the full set of 150 secrets to compute Exposure and Reducible Exposure, but this set of 48 "measurable" secrets is the focus of our more detailed analysis in Figure 4.

4.1.4 Inference Models. We chose a diverse set of three inductive KGC models that each operate quite differently, none of which were exposed to our secrets during training. To enforce that, we train each model using only the public graph and then perform inference on the public graph plus the review set. It is our use of inductive models that makes such a process possible.

Our first inference model, RuleN, is a simple statistical model [29]. RuleN samples the occurrences of each relation type and observes the frequency with which different relation sequences (called "rules") are found between the subjects and objects of this relation type. Like many inductive approaches, RuleN only considers the relations along a path, not the nodes through which a path passes. The score for RuleN's prediction is simply the frequency at which the highest-scoring rule present in the sample was observed. This rule-based approach makes it straightforward to identify RuleN's MDR, since only relations along paths that contain the highest-scoring rule can impact RuleN.

Graph Inductive Learning (GraIL) [44] is our second model. GraIL is an inductive model focused on the relations surrounding the secret. It trains a graph neural network to characterize neighborhoods in the graph. This approach can potentially capture far more nuance than RuleN, although it requires extensive computation to reason over many possible relation types, a problem when training on large graphs. Moreover, GraIL does not identify a specific path on which the MDR would be found as a byproduct of its reasoning, so we exhaustively check the effect on GraIL's inference of each possible single-relation redaction in a three-hop neighborhood. The relation that produces the largest reduction in Exposure is selected as GraIL's MDR.

Our third model is the Neural Bellman-Ford Network [52], which we call "Bellman-Ford" for short. Based on a generalized Bellman-Ford algorithm for identifying shortest paths to any point in a graph, this model learns ways in which paths can be represented and combined. Like RuleN and GraIL, Bellman-Ford considers only relation paths, and is thus an inductive model. However, it is faster to train than GraIL, and Zhu et al.[52] have also suggested an approach for estimating the relative impact of each path on its overall decision, potentially simplifying our search for the MDR.

In our result tables, we refer to these models as RN (RuleN), Gr (GraIL), and BF (Bellman-Ford). We restrict the neighborhood considered by each model to three hops in our experiments, which has been reported as a suitable choice for each when performing the KGC inference task on this collection [44, 52]. We consider each nomination model alone and in a two-model ensemble. When ensembling models for nomination, we use reciprocal rank fusion with the standard parameter (60) to merge results from each model [15]. We then select the relation with the largest reciprocal rank after reciprocal rank fusion as the joint nomination. Despite the potential for two of our models (RuleN and Bellman-Ford) to constrain the search space, we use exhaustive search of all potential single-relation redactions in the three-hop neighborhood of the secret relation with every model. This improves comparability by avoiding confounds from errors in Bellman-Ford's path impact estimates.

4.2 Results

Table 2 and Figure 4 show our results.

4.2.1 Evaluators. Figure 4 shows per-secret Reducible Exposure for each of the 48 secrets that have Reducible Exposure for at least one of our evaluator models. Among those evaluators, RuleN experiences the largest number of changes, followed by GraIL, and then Bellman-Ford in a distant third place. However, the magnitude of the changes exhibits the inverse behavior: while Bellman-Ford sees fewer changes in rank from a single-relation redaction, it sees the three most extreme changes, with GraIL seeing the next largest change magnitudes. This pattern matches the reported relative inference ability of the models on the KGC inference task: Bellman-Ford outperforms GraIL [52], and GraIL outperforms RuleN [44]. As expected (since those comparisons were also made using negative sampling), this pattern can also be seen in the Exposure measures in Table 2. This correlation makes sense, since large reductions in reciprocal rank are only possible when the secret being protected was very highly ranked to begin with.

Another observation from Figure 4 is that each evaluator has some secrets on which only it experiences rank changes. Even where overlap occurs, Reducible Exposure varies substantially between evaluators. We also see differences between models on secrets with more or less common relation types. The 48 secret relations in Figure 4 are sorted from most common type (*film release region*) to least common type (*music track contribution role*). Reducible Exposure for RuleN exhibits little dependence on how common or rare the relation type is, but Bellman-Ford exhibits a clear preference for mid-range relation types (i.e., those with higher entropy).

The left side of Table 2 shows evaluation model results. We see that the strongest evaluator is Bellman-Ford, because it achieves the highest Exposure and the lowest Reducible Exposure. We can also see the impact of RuleN's smaller but more numerous reductions in reciprocal rank in its greater Reducible Exposure than either other evaluator. The larger Reducible Exposure for RuleN indicates its inference can be affected by redaction to a greater degree (on average) than that of GraIL, and the same analysis indicates GraIL's inference can be affected more than Bellman-Ford's.

4.2.2 Cross-Model Nomination. As the right side of Table 2 shows, RuleN seems to be the best other nominator against both our strongest evaluator (Bellman-Ford) and against GraIL. When the evaluator is the same, we can conduct a paired t-test for statistical significance. Doing so, we find that RuleN is significantly better than Bellman-Ford as a nominator (when GraIL is the evaluator), but that RuleN is statistically indistinguishable from GraIL when Bellman-Ford is the evaluator. This lack of significance may result from the small number of samples (17) when Bellman-Ford is the evaluator, however. RuleN's impressive nomination performance against other evaluators results from one simple fact – if a nominator sees no difference for a secret, it can't make a redaction nomination. As Figure 4 shows, RuleN chooses some relation to redact for 26% more secrets as GraIL, and more than twice as many secrets as for Bellman-Ford. When doing cross-model nomination, we see that breadth beats depth. This premium on breadth suggests that using a nominator ensemble might be worth exploring. As the far right column in Table 2 shows, ensembling the other two (non-evaluator) models using reciprocal rank fusion was never better than the better of the two models being combined. There may, however, still be a case to be made for ensembles, particularly in settings without a single best nominator, as we observed in Section 3.

4.3 Preferential Attachment for Secret Selection

In practice, we expect secrets to exhibit a degree of locality because some topics, such as plans for future monetary policy, may involve many secrets. We can model that locality using preferential attachment [4, 24]. To explore the impact of differences in the distribution of secret relations, we selected 150 new secrets using a greedy method in which half the time we select a relation that shares a subject or an object with at least one other secret; the other half of the time we select a relation at random. When selecting a relation with a shared subject or object, the selection probability is the count of previously selected secret relations sharing that node. This approach encourages large clusters of secrets to develop.

As Table 3 shows, this changes things quite a bit. Effects on Exposure vary, with two models (RuleN and GraIL) seeing increases, suggesting this is an easier condition for those models. All models saw an increase in Reducible Exposure, indicating redactions can be more effective in such cases, and there are now substantially more secrets (72) for which redaction can affect at least one of the models. GraIL is now the best evaluator, and it is also the best cross-model nominator against both of the other evaluator models.

While work remains to be done on realistically modeling the distributions of real secrets (which will surely vary), these experiments make it clear that how secrets are distributed in a real collection can be consequential for both nominator and evaluator performance. One approach would be to perform experiments on information that is actually secret, but reproducing such results might require waiting decades for the test set to be released! Better would be to first study the Swiss-cheese distributions of actual secrets, and then to share the recipe for replicating that Swiss-cheese pattern in other collections. With such test sets, we could focus our work on experiments closer to the real problem.

4.4 Inference on Larger Graphs

While our experiments to this point have focused on the ability of a nominated redaction from one model to make inference more difficult, we also need to consider scalability to larger knowledge graphs. Some models, like RuleN, are designed to employ sampling and may (with progressively lower sampling rates) thus scale to large graphs. Other models, like GraIL, lack sampling in their baseline implementation and struggle with larger graphs (GraIL has an estimated run time on the full FB15k-237 dataset of over a month using a V100 GPU [52]). Exploring larger and more complex graphs will require more efficient approaches to inference. Our redaction task has a key difference from KGC that we can exploit to this end: we only care about our small number of secret relations in the graph, and these are potentially constrained to only a few relation types. As a result, we only need models that can predict specific relation types. If we focus the training of models on a small number of relation types, we can substantially reduce the training required and thus improve the scalability of these models.

To test this hypothesis, we focus on our least efficient model, GraIL. Since the scalability bottleneck is training the inference model, we return to the original KGC task on the full V1 (and later V4) FK15k-237 data rather than our public/review/secret splits. This approach provides more training and testing data, and enables straightforward transitions to the other FB15k-237 divisions, However, we choose the relation types on which we train and test

Table 2: Model effectiveness for 150 secret relations chosen randomly, with exhaustive search for relation redaction. RN: RuleN, Gr: GraIL, BF: Bellman-Ford.

	Total	Reducible	Nominator's Fraction↑			
Evaluator	Exposure↑	Exposure↓	RN	Gr	BF	Both
RN	0.211	0.041	–	**77.8%**	67.0%	74.5%
Gr	0.372	0.028	**57.2%**	–	40.1%	51.8%
BF	0.395	0.012	**77.1%**	53.2%	–	53.3%

Table 3: Model effectiveness for 150 secret relations chosen with preferential attachment; exhaustive relation redaction.

	Total	Reducible	Nominator's Fraction↑		
Evaluator	Exposure↑	Exposure↓	RN	Gr	BF
RN	0.228	0.053	–	**56.1%**	48.1%
Gr	0.403	0.037	**48.3%**	–	46.3%
BF	0.305	0.018	26.0%	**56.8%**	–

Table 4: Improvement in training time for Focused Grail.

Dataset	Relations	Model	Total Exposure↑	Training Time↓
V1	4,245	GraIL	**0.659**	218.70 minutes
		Focused GraIL	0.581	**5.41** minutes
V4	27,203	GraIL	**0.649**	7,747.62 minutes
		Focused GraIL	0.640	**20.64** minutes

from among those present in our V1 secrets. We modified GraIL's subgraph creation to only consider relations of one type. Since each secret has just one relation type, this design matches our one-secret-at-a-time evaluation approach. We selected 22 of the 26 relation types of our randomly selected secrets in Section 4.2 from the V1 split (omitting 4 types not present in the V1 test set), and we subsampled 17 relation types at random from those secrets for use in V4.[5] For each relation type, we trained the model on FB15k-237 V1 using only the 3 hop subgraphs around relations of only that specific type. We then performed the inference task, restricting the cases on which it was tested to test set triples containing our target relation type. We call this system *Focused GraIL*. Our Exposure measure remains useful in this setting, but it as a macroaverage, first by relation and then by relation type, so more and less common relation types contribute equally.

As expected, training time scales linearly with the proportion of the graph represented by a secret's relation type, with our tested relation types averaging a respective 2.1% and 1.9% of the V1 and V4 splits on a GTX 1080, with 8GB of VRAM, although there is some overhead in the initial processing of the graph. In practical applications, we might expect sets of secrets to include more than one relation type. Our approach could be applied in such cases by training one model per relation type or by training on just the subset of the relation types from the full collection that span the secrets that require protection. The timing data in Table 4 shows training all 22 Focused models on V1 would be nearly twice as fast as full GraIL training, although training the 180 modified models needed to cover all relation types in V1 would take substantially longer than full GraIL training. This approach is thus most useful when the types of relations under consideration are limited. The measured Exposure differences between GraIL and Focused GraIL are not statistically significant by a two-sided paired t-test for V1 or V4, with about as many wins as losses across relation types. Of the 22 relation types in V1, reciprocal rank increased for 9 and decreased for 11 (2 saw no change). Moreover, on the larger V4 set, the measured difference in the mean across 17 relation types is quite small. There are several possible sources of variation, including the simpler learning task in Focused GraIL (one vs. many rather than many vs. many) and changes to the distribution of positive and negative training subgraphs. To check for systematic differences, we performed redaction nomination experiments using five randomly selected V1 secrets, comparing each redaction nominated by the Focused GraIL model to the redaction nominated by the full GraIL model. We observed identical nominations for each of these secrets.

4.5 Overall Remarks

Our knowledge graph experiments support the usefulness of Reducible Exposure as an evaluation measure, albeit with different details to leverage more nuanced measurements of rank among confounds rather than answer correctness, and we found our nominator/evaluator framework could be usefully applied without modification. As expected, the more highly structured knowledge graph setting also offered greater scope for additional experimentation, including on system combination, secret selection, and scalability.

5 Limitations

First, we note that there are no guarantees beyond the specific models we have tested, for both text and knowledge graphs. While we saw strong performance of some models as nominators and others as potentially strong evaluators, other models, and in particular models that don't yet exist, could change these results. It is largely this challenge that has discouraged experimental work on detecting mosaicing risks before now—no matter what we do, we still won't know what we don't know. While that remains true, we note that we regularly use models that are imperfect for a number of tasks, including, for example, search. Review backlogs are large and growing, in part because the review process is expensive, and in part because some inference risks could be highly consequential. In such settings, even imperfect solutions may well be useful. Additionally, our focus on single-sentence or single-relation redaction has simplified the real task, in which several redactions might be needed to prevent an inference. Of course, single-redaction systems could be cascaded in a greedy manner to address this challenge, but we leave the more general (and more computationally complex) task of finding optimal redaction sets to future work. Finally, we note that our knowledge graph evaluation has focused on identifying which models are better or worse, both as nominators and as evaluators, and not yet on where to draw the line about how much inference risk is acceptable. This is fundamentally a policy question since, while preserving secrets is important, the ultimate goal of the review process is to release what can safely be released. Otherwise, we would simply never release anything. We can, however, inform the policy development process by creating evaluation measures that can help to characterize how much assistance an automated system can actually provide to a human reviewer. For that, we will need to move from ranked measures of relative risk to something closer to an actual risk-reward tradeoff framework.

6 Conclusion

We have introduced a method for reducing the likelihood of inference, adapting existing inference models to protect secrets by nominating information for redaction, and characterizing the risks to secrets through a flexible Exposure and Exposure reduction framework. We have done this in both same-model (i.e., same nominator and evaluator) and cross-model settings, in both text and knowledge graphs. We found that there is no one model that rules them all—sometimes the strongest evaluator is not the best choice as a nominator (when the evaluator model differs). We also found that the distribution of secrets has a substantial impact on inference model performance. Finally, we developed a method to adapt models to exploit aspects of our problem that enable them to overcome existing limitations in handling large data. Our results indicate this is a promising area of study, with considerable opportunities for further research.

[5] Subsampling was done for practical reasons, since full GraIL is very slow on V4.

References

[1] Johnathan Abel. 2014. Do You Have to Keep the Government's Secrets? Retroactively Classified Documents, the First Amendment, and the Power to Make Secrets out of the Public Record. *University of Pennsylvania Law Review* 163 (2014), 1037–1097.

[2] Asma Alnemari, Rajendra K Raj, Carol J Romanowski, and Sumita Mishra. 2019. Protecting Personally Identifiable Information (PII) in Critical Infrastructure Data using Differential Privacy. In *2019 IEEE International Symposium on Technologies for Homeland Security (HST)*. IEEE, 1–6.

[3] Leif Azzopardi, Emma Nicol, Jo Briggs, Wendy Moncus, Burkhard Schafer, Callum Nash, and Melissa Duheric. 2025. Assessing Risks in Online Information Sharing. *Conference on Human Information Interaction and Retrieval (CHIIR)* (2025).

[4] Albert-László Barabási and Réka Albert. 1999. Emergence of Scaling in Random Networks. *Science* 286, 5439 (1999), 509–512.

[5] Peru Bhardwaj, John Kelleher, Luca Costabello, and Declan O'Sullivan. 2021. Poisoning Knowledge Graph Embeddings via Relation Inference Patterns. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, Online, 1875–1888. https://doi.org/10.18653/v1/2021.acl-long.147

[6] Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating Embeddings for Modeling Multi-Relational Data. *Advances in Neural Information Processing Systems* 26 (2013).

[7] Robert Bredereck, Vincent Froese, Sepp Hartung, André Nichterlein, Rolf Niedermeier, and Nimrod Talmon. 2015. The Complexity of Degree Anonymization by Vertex Addition. *Theoretical Computer Science* 607 (2015), 16–34.

[8] Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. 2017. k-Degree Anonymity and Edge Selection: Improving Data Utility in Large Networks. *Knowledge and Information Systems* 50 (2017), 447–474.

[9] Jordi Casas-Roma, Jordi Herrera-Joancomartí, and Vicenç Torra. 2017. A Survey of Graph-Modification Techniques for Privacy-Preserving on Networks. *Artificial Intelligence Review* 47 (03 2017). https://doi.org/10.1007/s10462-016-9484-8

[10] Jeffrey A. Charleston. 2023. Improving Declassification: Applying Machine Learning to Diplomatic Cable Review. *Perspectives on History* (2023). https://www.historians.org/perspectives-article/improving-declassification-applying-machine-learning-to-diplomatic-cable-review-october-2023/

[11] Xihui Chen, Sjouke Mauw, and Yunior Ramírez-Cruz. 2019. Publishing Community-Preserving Attributed Social Graphs with a Differential Privacy Guarantee. *Proceedings on Privacy Enhancing Technologies* 2020 (2019), 131 – 152. https://api.semanticscholar.org/CorpusID:202540124

[12] Zhe Chen, Yuehan Wang, Bin Zhao, Jing Cheng, Xin Zhao, and Zongtao Duan. 2020. Knowledge Graph Completion: A Review. *IEEE Access* 8 (2020), 192435–192456.

[13] Shiqi Cheng, Xuefei Zhang, Yao Sun, Qimei Cui, and Xiaofeng Tao. 2024. Knowledge Discrepancy Oriented Privacy Preserving for Semantic Communication. *IEEE Transactions on Vehicular Technology* (2024).

[14] Christopher Clark and Matt Gardner. 2018. Simple and Effective Multi-Paragraph Reading Comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 845–855.

[15] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. Reciprocal Rank Fusion Outperforms Condorcet and Individual Rank Learning Methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 758–759.

[16] Ming Ding, Chang Zhou, Qibin Chen, Hongxia Yang, and Jie Tang. 2019. Cognitive Graph for Multi-Hop Reading Comprehension at Scale. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 2694–2703. https://doi.org/10.18653/v1/P19-1259

[17] Yuwei Fang, Siqi Sun, Zhe Gan, Rohit Pillai, Shuohang Wang, and Jingjing Liu. 2020. Hierarchical Graph Network for Multi-hop Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8823–8838.

[18] Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-Contrastive Estimation of Unnormalized Statistical Models, with Applications to Natural Image Statistics. *Journal of Machine Learning Research* 13, 2 (2012).

[19] Michael Hay, Gerome Miklau, David Jensen, Don Towsley, and Philipp Weis. 2008. Resisting Structural Re-Identification in Anonymized Social Networks. *Proceedings of the VLDB Endowment* 1, 1 (2008), 102–114.

[20] Fred Hiatt. 1984. U.S. Nuclear Ship Uses Canal. *The Washington Post* (6 November 1984). https://www.washingtonpost.com/archive/politics/1984/11/06/us-nuclear-ship-uses-canal/4df3286a-62cc-4bf3-8599-89ed83bd31de/

[21] Anh-Tu Hoang, Barbara Carminati, and Elena Ferrari. 2023. Protecting Privacy in Knowledge Graphs with Personalized Anonymization. *IEEE Transactions on Dependable and Secure Computing* (2023).

[22] Haiping Huang, Dongjun Zhang, Fu Xiao, Kai Wang, Jiateng Gu, and Ruchuan Wang. 2020. Privacy-Preserving Approach PBCN in Social Network With Differential Privacy. *IEEE Transactions on Network and Service Management* 17, 2 (2020), 931–945.

[23] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714* (2023).

[24] Jay Lee, Manzil Zaheer, Stephan Günnemann, and Alex Smola. 2015. Preferential Attachment in Graphs with Affinities. In *Artificial Intelligence and Statistics*. PMLR, 571–580.

[25] Ruosen Li and Xinya Du. 2023. Leveraging Structured Information for Explainable Multi-hop Question Answering and Reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 6779–6789. https://doi.org/10.18653/v1/2023.findings-emnlp.452

[26] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. 2007. L-Diversity: Privacy Beyond k-Anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1, 1 (2007), 1–52.

[27] Vaibhav Mavi, Anubhav Jangra, and Adam Jatowt. 2024. Multi-hop Question Answering. *Foundations and Trends in Information Retrieval* 17, 5 (2024), 457–586.

[28] Graham McDonald. 2019. *A Framework for Technology-Assisted Sensitivity Review: Using sensitivity classification to prioritise documents for review*. Ph. D. Dissertation. University of Glasgow.

[29] Christian Meilicke, Manuel Fink, Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, and Heiner Stuckenschmidt. 2018. Fine-Grained Evaluation of Rule- and Embedding-Based Systems for Knowledge Graph Completion. In *International Workshop on the Semantic Web*.

[30] Andriy Mnih and Yee Teh. 2012. A Fast and Simple Algorithm for Training Neural Probabilistic Language Models. *Proceedings of the 29th International Conference on Machine Learning, ICML 2012* 2 (06 2012).

[31] Hitarth Narvala, Graham McDonald, and Iadh Ounis. 2024. Displaying Evolving Events via Hierarchical Information Threads for Sensitivity Review. In *European Conference on Information Retrieval*. Springer, 261–266.

[32] George Washington University National Security Archive. 2019. Redactions: The declassified file. Website https://nsarchive.gwu.edu/briefing-book/foia/2019-04-18/redactions-declassified-file, visited July 7, 2024.

[33] Alexandra Olteanu, Jean Garcia-Gathright, Maarten de Rijke, Michael D Ekstrand, Adam Roegiest, Aldo Lipani, Alex Beutel, Alexandra Olteanu, Ana Lucic, Ana-Andreea Stoica, et al. 2021. FACTS-IR: Fairness, Accountability, Confidentiality, Transparency, and Safety in Information Retrieval. In *ACM SIGIR Forum*, Vol. 53. ACM New York, NY, USA, 20–43.

[34] OpenAI. 2024. Models. https://platform.openai.com/docs/models. visited January 21, 2025.

[35] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2019. Investigating Robustness and Interpretability of Link Prediction via Adversarial Modifications. In *Proceedings of NAACL-HLT*. 3336–3347.

[36] Pouya Pezeshkpour, Yifan Tian, and Sameer Singh. 2020. Revisiting Evaluation of Knowledge Base Completion Models. In *Proceedings of the Conference on Automated Knowledge Base Construction (AKBC)*.

[37] David E Pozen. 2005. The Mosaic Theory, National Security, and the Freedom of Information Act. *The Yale Law Journal* (2005), 628–679.

[38] Paul S. Prueitt. 1999. Similarity Analysis and the Mosaic Effect. In *Proceedings of the 1999 Symposium on Document Image Understanding Technology*.

[39] Jianwei Qian, Xiang-Yang Li, Chunhong Zhang, Linlin Chen, Taeho Jung, and Junze Han. 2019. Social Network De-Anonymization and Privacy Inference with Knowledge Graph Model. *IEEE Transactions on Dependable and Secure Computing* 16, 4 (2019), 679–692. https://doi.org/10.1109/TDSC.2017.2697854

[40] Mahmoud F Sayed. 2021. *Search Among Sensitive Content*. Ph. D. Dissertation. University of Maryland, College Park.

[41] Larry Spitz. 1997. Duplicate Document Detection. In *Document Recognition IV*, Vol. 3027. SPIE, 88–94.

[42] Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. 2019. RotatE: Knowledge Graph Embedding by Relational Rotation in Complex Space. In *Proceedings of the Seventh International Conference on Learning Representations (ICLR)*.

[43] Latanya Sweeney. 2002. k-Anonymity: A Model for Protecting Privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10, 05 (2002), 557–570.

[44] Komal K. Teru, Etienne G. Denis, and William L. Hamilton. 2020. Inductive Relation Prediction by Subgraph Reasoning. In *Proceedings of the 37th International Conference on Machine Learning (ICML'20)*. JMLR.org, Article 876, 10 pages.

[45] Kristina Toutanova and Danqi Chen. 2015. Observed Versus Latent Features for Knowledge Base and Text Inference. In *Proceedings of the 3rd Workshop on Continuous Vector Space Models and their Compositionality*. Association for Computational Linguistics, Beijing, China, 57–66. https://doi.org/10.18653/v1/W15-4007

[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal

Azhar, et al. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971* (2023).

[47] Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge Graph Embedding by Translating on Hyperplanes. *Proceedings of the AAAI Conference on Artificial Intelligence* 28, 1 (Jun. 2014). https://doi.org/10.1609/aaai.v28i1.8870

[48] Bishan Yang, Wen tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2014. Embedding Entities and Relations for Learning and Inference in Knowledge Bases. In *International Conference on Learning Representations*. https://doi.org/10.48550/arXiv.1412.6575

[49] Hui Yang, Ian Soboroff, Li Xiong, Charles L.A. Clarke, and Simson L. Garfinkel. 2016. Privacy-Preserving IR 2016: Differential Privacy, Search, and Social Media. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) *(SIGIR '16)*. Association for Computing Machinery, New York, NY, USA, 1247–1248. https://doi.org/10.1145/2911451.2917763

[50] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (Eds.). Association for Computational Linguistics, Brussels, Belgium, 2369–2380. https://doi.org/10.18653/v1/D18-1259

[51] Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Liu Yong, and Shen Huang. 2024. End-to-End Beam Retrieval for Multi-Hop Question Answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 1718–1731.

[52] Zhaocheng Zhu, Zuobai Zhang, Louis-Pascal Xhonneux, and Jian Tang. 2021. Neural Bellman-Ford Networks: A General Graph Neural Network Framework for Link Prediction. In *Neural Information Processing Systems*.