What If Alice Wants Her Story Told?

Anindya Bhandari* Allison Bishop†

Abstract

In this paper, we pose the problem of defining technical privacy guarantees for anonymizing stories. This is a challenge that arises in practice in contexts like journalism and memoir writing, and can be crucial in determining whether a story gets told and how effectively it is presented. While recent decades have seen leaps and bounds in the development of approaches like differential privacy for numerical and categorical data sets, none of these techniques are directly applicable to settings where narrative structure is crucial to preserve. Here we begin an exploration of what kind of definitions could be achievable in such settings, and discuss the building blocks and interdisciplinary collaborations that could shape new solutions. Having scientific guarantees for anonymity in narrative forms could have far-reaching effects - in the peace of mind of potential tellers and subjects of stories, as well as in the legal sphere. This could ultimately expand the kinds of stories that can be told.

1 Introduction

A cryptography paper often starts with a story. Alice has a goal: she wants to communicate secretly with Bob, or she wants to delegate a difficult computation to Charlie, etc. Alice also faces an obstacle: the communication lines are tapped by Eve! Charlie cannot be trusted! But not to fear - by the end of the introduction, Alice has found a solution. A new cryptographic protocol has emerged to save the day, allowing Alice and the other residents of Cryptoville to live happily ever after! Or not. (Sometimes it's an impossibility result.)

These stories we tell as cryptographers are not strictly necessary. We could, as mathematicians often do, skip the pleasantries and step straight into the technicalities of our creations. But presumably we tell stories because they help our readers internalize the meaning of what we are doing and imagine its potential applications. Stories have explanatory power that can be hard to achieve in any other form.

Beyond mere explanation, written stories also take readers through emotional arcs that can enable them to connect viscerally with a character's experiences. This emotional connection is often crucial to fulfilling the storyteller's purpose, and can be a uniquely powerful tool in fostering empathy across disparate experiences. Abstractions or statistics simply fail to do this to the same extent.

But this same power can be dangerous to the subjects of a story, or the storyteller herself. Identifiable villains in a story may become targets for harassment in real life, and storytellers may risk legal, social, and physical consequences when a subject of a story doesn't like how they were portrayed. Such concerns are often handled today by removing names or other clearly identifying details from non-fiction stories. In a journalistic context, protection of sources is a core tenet that allows a wider set of stories to told. In memoir or other cases where a public-facing subject is willing to reveal their own identity, other names and details may be changed to protect the privacy of others, or to protect the named subject from potential repercussions.

*Hidden Road

†Proof Trading and City College, CUNY

It may be hard to gain confidence that such measures will be protection enough. Sometimes this may force would-be storytellers back into the safer realms of dry statistics and abstractions, to the detriment of their intended messages. How can Alice anonymize, for example, a personal story about domestic abuse? If she writes the words "I am a domestic abuse survivor" in any context that can be traced back to her, a former partner may infer that Alice is talking about them, and Alice may expose herself to devastating personal consequences. Certainly this threat may have a chilling effect on Alice, and she may struggle with the decision to tell her story in any form at all.

It is not difficult to imagine that the stigma attached to domestic abuse feeds off this stifling effect. We may be told sobering statistics like those maintained for the US by the National Coalition Against Domestic Violence [11], establishing that roughly one in four women and one in nine men in the US experience severe intimate partner violence or stalking in their lifetimes. But as shocking as such statistics can be, they may not be as effective in fostering empathy and rallying support as personal stories. Statistics alone are easy for people to distance themselves from ("no one I know") or perversely, to draw a nearly opposite conclusion from ("since it's common, it must not be that bad").

Alice's story could be an immensely powerful thing. For audiences uninitiated in the experience of domestic abuse, Alice's story could represent an opportunity to connect emotionally with a survivor's experience and gain a new level of understanding, lessening stigma and opening new possibilities for supportive engagement. For other survivors, Alice's story may help them to feel less alone and help them on a path to healing. For Alice herself, knowing her story has been told may help provide closure and purpose.

Alice is often left to balance these benefits with the risks on her own. Counselors, therapists, social workers, and lawyers may lend an experienced ear, but not every survivor will gain access to such resources. And even with the benefit of expert advice, Alice may be embarking on a fraught and dangerous exercise. How can she decide that a given version of her story is safe to tell, even if it's published through a medium that cannot be directly traced back to her?

Obviously this is not a wholly or even primarily scientific problem, and we should not delude ourselves by trying to fully reduce it to one. However, there is a small and meaningful part of this that could be helped by scientific development. For data in numerical or categorical form (rather than narrative), the last few decades have seen rapid evolution of privacy definitions and privacy-preserving tools. It may be possible to draw inspiration from this and develop new privacy definitions and techniques for use in a narrative context.

1.1 The Evolution of Privacy Tools for Numerical and Categorical Data

The landscape of privacy definitions and tools in the setting of numerical and categorical data began in a similar state to what happens for narratives today, with basic removal of "Personally Identifying Information" (PII) serving as a standard. However, demonstrable attacks revealed troubling limitations of this approach. Data fields like gender, zip code, and birth date that are not considered personally identifying on their own can nonetheless be used in combination to link subjects across different data sets and often de-anonymize them [16].

Sweeney's landmark paper in 2002 [16] introduced the concept of k-anonymity, a stronger standard of privacy than the mere removal of PII. To achieve k-anonymity, a set of data fields like gender, zip code, etc. must be flagged as *quasi-identifiers*, and data must be processed to ensure that every combination of values for the quasi-identifiers corresponds to at least k data points. For example, suppose that each data point represents a person and we declare gender and age as our only quasi-identifiers and set the parameter $k = 3$. If every gender and age combination that appears in our data is shared by at least three people, then we satisfy

k-anonymity for these parameters. If our data does not have this property, then we may need to do things like rounding the age values more coarsely in order to satisfy our requirements.

Intuitively, k-anonymity tries to capture a "safety in numbers" effect. But there are some things that can still go wrong. For one, safety in numbers will not protect from inference if, say, all of the $\geq k$ people share a common value for a sensitive data field. For example, let's consider a medical data set where there are a host of quasi-identifiers, such as an individual's demographic data. Along with that, there exists a set of non-quasi-identifier fields, such as sensitive medical information. Let's suppose that one of these fields is a boolean expressing whether or not the individual has lymphoma. From the perspective of an adversary who knows a target individual's quasi-identifiers only, the target individual is indistinguishable from those $\geq k-1$ other individuals with whom they share a set of quasi-identifiers. But if the entire group of $\geq k$ all have that boolean as $True$, the adversary can learn that the individual has lymphoma!

In addition to the above problem, the more fields that are labeled as quasi-identifiers, the more difficult k-anonymity is to achieve, and the more utility is lost. And yet, even seemingly innocuous things can turn out to be quasi-identifiers. An infamous example is the de-anonymization attack on the Netflix data set [12] which linked "anonymized" movie ratings to public IMDB profiles.

The concept of differential privacy was introduced in 2006 [7]. Differential privacy centers on a stronger goal than safety in numbers, aiming instead to give data contributors a guarantee like: "the distribution of any outcome in a world with your data included is statistically close to its distribution in a world without your data included." This kind of guarantee is only achievable if the collector and processor of data can be trusted to add randomness along the way to a final output, enough to provide ample cover for any individual's impact on the result. This kind of guarantee should hold *even in the presence of arbitrary auxiliary information*, meaning that we are making no assumptions about what an adversary might know about the data subjects or what outside data sources they may try to link to.

It's clearest to imagine how differential privacy can be achieved in the context of simple statistical queries. Let's imagine, for example, that we want to conduct a survey and release an aggregate count of how many respondents share a certain property. In this case, any one person's answer affects the count by $+0$ or $+1$. If we publish the exact count, a hypothetical adversary who knew all of the true answers except one could infer the missing response. In isolation, this may seem too extreme to be a reasonable concern. But trying to draw firm lines between an all-knowing adversary and a "knows everything reasonable" adversary is a shaky path, and may not be necessary in many settings. For this counting example, we can add noise to the total count so that even the all-knowing adversary cannot gain much of an advantage in inferring the missing response. A published answer of "100," for instance, could be explained by a real count of 100 and a noise value of 0, or a real count of 99 and a noise value of 1, or a real count of 101 and a noise value of -1, etc. If noise values that are ± 1 away from each other have similar enough probabilities, a survey respondent can be confident that the published answer would have had a similar probability of occurring without their input. This presumably lowers any disincentives to participate.

Nearly two decades of academic research have built upon, expanded, and adapted this basic premise (see e.g. [4] for a recent survey). Practical implementations and deployments are emerging, including the use of differential privacy for the US Census [1]. Several features of differential privacy make it a strong foundation to build upon: 1. intuitive metrics for utility (e.g. how close to the real answer is the released answer in expectation), 2. intuitive privacy guarantees, even against arbitrarily knowledgeable adversaries (the same outcomes would have happened with similar probability without your participation), and 3. customizability (variants

of differential privacy can be fit to a variety of situations and desired utility/privacy trade-offs).

The privacy goal in differential privacy, that any one data point should not overly influence the distribution of the output, can even be re-framed as a utility goal if we want to perform statistical calculations that are robust to outliers. In fact, connections between differential privacy and robust statistics have been formalized in works like [6, 10]. Such an interpretation stands in sharp contrast to a popular complaint about differential privacy, that its addition of randomness represents an inherently undesirable form of inaccuracy [9].

1.2 An Analogous Evolution for Narratives?

There are some strong parallels between the world of narratives and the world of numerical/categorical data. In both cases, the abundance of information sources outside of the source to be anonymized and the insider knowledge that an adversary may have about a particular target represent a challenge. Basic techniques like the removal of names and other obviously identifying information may not be sufficient.

Also in both settings, there are plenty of things we want to learn that do not feel inherently tied to the specifics of any one story or data point. In the quantitative setting, the advancing field of robust statistics (see e.g. [5]) shows that many valuable analyses can be performed on data in a way that does not depend heavily on the exact values of a small fraction of the data points. In the narrative context, we often extract meaning that feels universal, even when our connection to the material is shaped by the specifics.

This combination of meaningful generalization and strong adversarial threat makes both setting ripe for a progression of scientific tools to help navigate the complex relationship between utility and privacy. But there are also some crucial differences between the settings. Stories in raw form are likely to be unique, in sharp contrast to individual numerical or categorical values. Stories also may draw heavily upon shared cultural context, and hence are rarely self-contained. Such differences do pose unique challenges in the narrative context. Nonetheless, we are hopeful that new techniques for balancing privacy and utility in the narrative context will be discovered, and we would like to provide a partial glimpse of what such techniques could look like. In the following sections, we will discuss potential structures for capturing utility in storytelling, potential privacy definitions, and gesture towards what a privacy-achieving story compiler could look like.

2 Capturing Utility

If we are going to discover new ways of preserving effective storytelling while obtaining more meaningful levels of privacy for story subjects, we will need to build a strong sense of what exactly we are trying to preserve. Naturally, since stories have been around a *long* time, there is a rich (dare we say overwhelming?) landscape of prior studies on how stories are typically constructed and what makes them tick. In the context of fiction, many works are focused on analyzing culturally relevant stories and tracking their commonalities and differences across time and space. Others are focused on tips for writers looking to breathe new life into tried-and-true structures. Both of these framings tend to coalesce on the same basic objects as units for analyses: plot, characters, and emotional arcs.

Plot For long forms of narrative fiction (e.g. novels, myths, movies), several common plot lines have been identified that underpin many seemingly disparate works, much like common chord progressions underlie many different popular songs. Perhaps the most famous of these is the "monomyth" or "hero's journey" described by Campbell in his 1949 book, "The Hero with

a Thousand Faces" [3]. This basic plot follows a hero who leaves the settings of ordinary life to have a fantastical adventure, battles the forces of evil, and ultimately returns home to reap rewards. A generalization of this has been identified as the sympathetic plot, which follows a goal-directed protagonist who confronts and conquers obstacles and thereby grows and achieves positive outcomes.

In "The Seven Basic Plots: Why We Tell Stories" [2], Booker lays out seven such common templates for fiction plots that seem to transcend culture, time, and style. These include "overcoming the monster," "rags to riches," "the quest," "voyage and return," "comedy," "tragedy," and "rebirth." All of these are plots that center on a hero who goes through an affecting series of experiences.

Similarly prescriptive structures can be found in screenwriting guides, such as the ubiquitous "Save the Cat! The Last Book on Screenwriting You'll Ever Need" by Snyder [14]. [Aside: amusingly, this book also has sequels.] Such guides are in fact *so* prescriptive that they caused one screenwriting teacher to quip, "the greatest trick the devil ever pulled was to write one mediocre movie and then convince the world to copy it."

Characters Some structural studies of stories center more on characters than plot. In "The Science of Storytelling," Storr [15] draws connections between storytelling and the operation of the human brain, centering characters rather than plot lines. This approach more deeply unites psychology and literary analysis.

Emotional Arcs A crucial part of the utility of Alice's story lives in the emotional arc that a reader experiences while reading it. This arc is generally a product of the reader connecting with one or more characters and mirroring their emotional arcs in the story. A recent work by Reagan, Mitchell, Kiley, Danforth, and Dodds [13] performs a analysis of emotional arcs in a collection of fiction stories, and finds that just six basic shapes dominate their structures. Much like the commonalities observed in plot lines, this provides some hopeful evidence that the core utility in stories does not rely heavily on unique attributes that would be nearly impossible to satisfyingly anonymize.

Transparency of Process We can view the writing of fiction as a strong way of protecting individual privacy while conveying the lessons of a personal story in an effective and emotionally resonant way. In this sense, Alice's problem could be declared "solved" by her writing a fiction novel very loosely inspired by her own emotional journey. But we do feel this loses something important. Which is not to say that Alice shouldn't write a novel - we're all for it! But a reader encountering a proclaimed work of fiction is in a fundamentally different mindset than a reader encountering a work of non-fiction, even when the text is the same. Some important goals that Alice may have for the process, such as combating stigma and feeling heard, may be considerably hindered by the medium of fiction as compared to non-fiction. In other potential application settings, such as journalism, simply shifting to fiction is a non-starter.

It is human nature to react to the same story differently when it is something that "happened" versus something that did not. How much a writer can stretch the truth without destroying the effectiveness of non-fiction is a tricky question, but one thing that will likely help here is transparency of process. If readers know what process was applied to get from an initial story to a more privacy-preserving version, they can have a fair basis for forming their own judgements. With fiction, there is no transparency of process.

Putting all of this together, we could imagine defining some kind of data structure to serve as a canonical form for the kind of story Alice wants to tell. This could hold elements deemed to be fundamental, like sets of characters, relationships between them, basic elements of plot,

emotional arcs, etc. We might hope that putting stories into such a canonical form would be a helpful step in identifying what about the story's original structure we want to preserve, and could also be done as a transparent process that storytellers and readers alike could intuitively understand.

3 Defining Security

We can try to imagine a progression of security definitions in the narrative context, similar to the one that exists for numerical and categorical data. Changing names of people and places, for example, is analogous to removal of PII. It may also be possible to define a notion of "quasi-identifying" details and impose some criterion analogous to k-anonymity across a set of stories, insisting that combinations of these details must appear in at least k stories. Similar to achieving k-anonymity in the numerical setting, details could be abstracted or redacted strategically to achieve a criterion like this.

Similar pitfalls of the k-anonymity setting for numerical data might exist here. Let's suppose, for example, that we have a large collection of stories about relationships. Some k of these might share a detail like celebrating a first anniversary at Eleven Madison Park, a world-famous restaurant in New York City. We might label this as a quasi-identifying detail. Now suppose that *all* such stories involve a miscarried pregnancy at some point. We might consider this as a sensitive but non-quasi-identifying detail, due to how common it is. Thus, an adversary who could hone in on the subset of stories due to the quasi-identifiers might infer a sensitive detail about a targeted individual who contributed her story.

A more ambitious security goal might be closer in spirit to differential privacy. On its face, this at first seems kind of weird. Can Alice's story really be effectively told in a way to could have existed *without* her? Well, probably. The core structure of Alice's story is unlikely to be particularly unique, and for each emotionally resonant detail, there may be reasonable equivalents that could be substituted. We might hope, then, for a trusted, randomized story processor that collects a large batch of stories, and then produces a set of outputs whose probabilities are quantitatively close under the removal or addition of any one input story.

This level of security may conflict with some desired utilities. Readers may not internalize the outputs as "real" enough with this level of disconnection between the outputs and individual inputs. And Alice may not feel the sort of closure that she is looking for, since none of the outputs are directly due to her contribution. This is the inherent flip side of a differential privacy-style guarantee: the same property that lowers disincentives to participation may similarly lower incentives to participation.

We might be able to find a kind of middle ground here by aiming for deniability more than output insensitivity. We could ask for a centralized processor that outputs a version of Alice's story that could have resulted from at least k other inputs from some set. To sketch out such a guarantee in a bit more detail, let's imagine our story processor SP as a randomized process that takes Alice's written story A as input and outputs a written story $\tilde{A}$. Let's imagine also that we have some reference set $\mathcal{S}$ of other stories (these could come from previously submitted stories, for example). We could let $\mathcal{S}_{\tilde{A}}$ denote the subset of stories $s \in \mathcal{S}$ such that SP would have produced $\tilde{A}$ from s with probability > 0. In other words,

$$\mathcal{S}_{\tilde{A}} := \{s \in \mathcal{S} \mid \mathbb{P}[SP(s) = \tilde{A}] > 0\}.$$

The probability here is over the randomness used by SP. If our probabilities may get arbitrarily small, we might set a threshold τ and instead define:

$$\mathcal{S}_{\tilde{A}} := \{s \in \mathcal{S} \mid \mathbb{P}[SP(s) = \tilde{A}] > \tau\}.$$

Our privacy definition would then require $|\mathcal{S}_{\tilde{A}}| \geq k$ for some specified set $\mathcal{S}$ and parameter k.

Similar to k-anonymity for numerical and categorical data, there is a threat from data pollution that may threaten the efficacy of this definition. If a malicious party can contribute content to $\mathcal{S}$, she can insert k stories of a similar form to Alice's, thereby making it seem "safe" for a perturbed version of Alice's story to be revealed, when it actually may not be. Whether this is a deal-breaking concern in a real scenario would heavily depend on *how* the set $\mathcal{S}$ is defined and collected. Even in a non-adversarial context, the common practice in machine learning of augmenting a data set is dangerous in our specific formulation, as we add no entropy to the data set while retaining the privacy guarantee $|\mathcal{S}_{\tilde{A}}| > k$, as augmented stories in this data set are still deterministically linked to their original constituents.

4 Construction Possibilities

If we were to choose a kind of "canonical form" for expressing core story elements such as plot, character, and emotional arcs, we could use this as a building block inside a privacy-preserving story compiler. The process could be roughly as pictured in Figure 1:

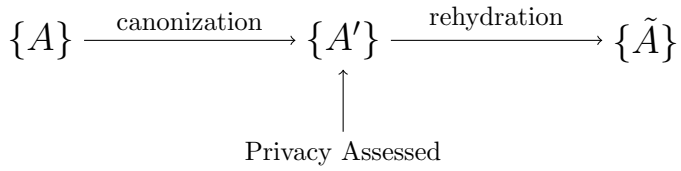

Figure 1: The privacy preserving canonization and rehydration process

We could start with a set of story texts, including the story text written by Alice, denoted by $\{A\}$. We could then put these into our chosen canonical form, yielding a set $\{A'\}$. At this stage, a chosen privacy definition could be assessed. For example, a k-anonymity style definition might involve a check that the same instantiation A' appears at least k times. A differential privacy style definition might involve taking the set of derived canonical forms and resampling it somehow to be less sensitive to individual inputs. After achieving the desired form of privacy at this layer, new story texts could be randomly sampled by a rehydration process, ultimately producing an output set $\{\tilde{A}\}$. Similar to numerical and categorical data, a k-anonymity style definition would allow a 1-to-1 mapping from inputs A to outputs $\tilde{A}$, while a differential privacy style definition would inherently lose this association.

The rehydration process would need to produce high quality and readable story texts, preserving the emotional resonance of the original texts to the greatest extent possible. Extensive user studies would need to be done to validate such properties. One might suspect that recent developments in large language models would be useful in designing a workable rehydration process. We would hope, however, to keep the rehydration component as *explainable* as possible, to reap the benefits of transparency of process in connecting the outputs to the inputs in the minds of readers. One nice feature of the workflow imagined in Figure 1 is that the rehydration process itself has no privacy requirements, but rather can be thought as analogous to post-processing of differentially private data, which can be done freely without invalidating the privacy guarantees (e.g. as discussed in [8]).

This basic schematic is not intended to be prescriptive. It is intended only to illustrate one potential path towards a solution.

5 Future Directions: An Interdisciplinary Approach

Admittedly, it is not common for a research paper to leave a basic task like "building the darn thing" as a future direction. But in this case, we believe that a proposed solution should likely involve deep interdisciplinary collaborations, and we think there is value in putting our initial thoughts out into the research community in a form that doesn't risk preempting diverse approaches.

We expect that literary scholars, journalists, psychologists, social workers, natural language processing experts, and legal/policy experts could all make very valuable contributions to the problem space we have been positing here. Problems around capturing the essential bones of story structure and detail would greatly benefit from the experience of those who study stories in a literary or cultural context, as well as those who extract first hand accounts from subjects in the practice of their work. Insights in psychology might prove crucial in developing and navigating trade-offs between privacy for storytellers and preserving "truth" for readers. Machine learning and privacy experts would be needed to process textual inputs, design and check privacy criteria, and help anticipate threats.

We hope that putting our thoughts down at this initial level might facilitate conversations with potential collaborators in these areas. Most crucially, potential users of any concrete application of ideas along these lines should play a shaping role in the development.

6 Acknowledgement

We are grateful for the contributions of an anonymous survivor of domestic abuse in helping to shape the motivation for this work.

References

[1] Abowd, J.M. and Hawes, M. B.: Confidentiality protection in the 2020 U.S. census of population and housing. Working Paper Number CED-WP-2022-003 (2022). https://www.census.gov/library/working-papers/2022/adrm/CED-WP-2022-003.html

[2] Booker, C.: The seven basic plots: Why we tell stories. Continuum (2006)

[3] Campbell, J.: The hero with a thousand faces. Pantheon Books (1949).

[4] Cummings, R., et. al. Challenges towards the next frontier in privacy. (2023). https://arxiv.org/abs/2304.06929

[5] Diakonikolas, I., and Kane, D.: Algorithmic robust statistics. Cambridge University Press (upcoming). https://sites.google.com/view/ars-book/home/

[6] Dwork, C., and Lei, J.: Differential privacy and robust statistics. STOC (2009)

[7] Dwork, C., McSherry, F., Nissim, K., and Smith, A.: Calibrating noise to sensitivity in private data analysis. Theory of Cryptography (2006)

[8] Dwork, C. and Roth, A.: The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science (2014)

[9] Lee, T. B.: Why the census invented nine fake people in one house. Slate (March 7, 2022). https://slate.com/technology/2022/03/privacy-census-fake-people

[10] Liu, X., Kong W., Kakade, S., and Oh, S.: Robust and differentially private mean estimation. NeurIPS (2021)

[11] National Coalition Against Domestic Violence https://ncadv.org/STATISTICS

[12] Narayanan, A. and Shmatikov, V.: Robust de-anonymization of large sparse datasets. IEEE Symposium on Security and Privacy (2008)

[13] Reagan, A.J, Mitchell, L., Kiley, D., Danforth, C.M., and Dodds, P.S.: The emotional arcs of stories are dominated by six basic shapes. EPJ Data Science **5**:31 (2016)

[14] Synder, B.: Save the cat!: The last book on screenwriting you'll ever need. Michael Wiese Productions (2005)

[15] Storr, W. The science of storytelling. William Collins (2019)

[16] Sweeney, L.: k-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness, and Knowledge-based Systems **10**:5 (2002)