No free delivery service
Epistemic limits of passive data collection in complex social systems

Maximilian Nickel
FAIR, Meta
New York, NY
maxn@meta.com

Abstract

Rapid model validation via the train-test paradigm has been a key driver for the breathtaking progress in machine learning and AI. However, modern AI systems often depend on a combination of tasks and data collection practices that violate all assumptions ensuring test validity. Yet, without rigorous model validation we cannot ensure the intended outcomes of deployed AI systems, including positive social impact, nor continue to advance AI research in a scientifically sound way. In this paper, I will show that for widely considered inference settings in complex social systems the train-test paradigm does not only lack a justification but is indeed invalid for any risk estimator, including counterfactual and causal estimators, with high probability. These formal impossibility results highlight a fundamental epistemic issue, i.e., that for key tasks in modern AI we cannot know whether models are valid under current data collection practices. Importantly, this includes variants of both recommender systems and reasoning via large language models, and neither naïve scaling nor limited benchmarks are suited to address this issue. I am illustrating these results via the widely used MOVIELENS benchmark and conclude by discussing the implications of these results for AI in social systems, including possible remedies such as participatory data curation and open science.

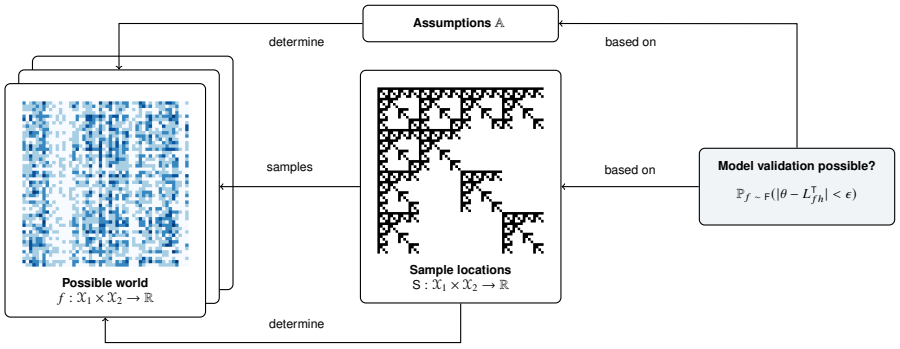


Figure 1: **Test validity in complex systems**. Given assumptions $\mathbb{A}$, target distribution T, data set $\mathcal{D} \sim \mathsf{S}^m$ from a sampling distribution S, and quality metric θ, an inference setting is *test-valid* if the difference between θ and the true risk L^{T}_{fh} can be bounded over the distribution of all possible worlds $f \sim \mathsf{F}$ consistent with $(\mathbb{A}, \mathcal{D})$.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Model validation, long taken to be "solved" via the train-test paradigm, has become one of the central challenges in modern machine learning and artificial intelligence. In unison with the dramatic increase of their capabilities, AI systems are now supposed to solve tasks of vastly expanded scope, including potentially AI-complete tasks such as open domain question answering, autonomous decision making, and ultimately, artificial general intelligence. Even before the recent triumphs of large language models and deep learning, Anderson [3] proclaimed "the end of theory" and the scientific method being obsolete due to the wonders of big data, large-scale computing, and data mining. At the same time, it is entirely unclear how to rigorously evaluate the quality of models for these ambitious tasks. This lack of proper evaluation can then materialize in persistent issues of deployed systems related to generalization, e.g., hallucination [30], out-of-distribution generalization [39], fairness [6], and generalization to the long-tail [23, 24]. Importantly, these issues do not only affect the accuracy of models in a vacuum, but can also affect their social impact if they are deployed in consequential social contexts [16]. In this paper, I aim to connect the former developments with the latter issues through the lens of epistemology. More concretely, I ask:

> **Research Question 1.** Given the ambitious tasks that we ask AI systems to solve and given how we currently collect data, *can we know* whether a model performs well for these tasks?

Answering RQ 1 positively is central not only for the deployment of machine learning systems, but also for scientific progress within artificial intelligence itself. After all, knowledge of a model's quality is a prerequisite to detect generalization issues and develop improved models. In deployed systems, a model's predictions are useless — as good as they might be — without knowing that they are, in fact, reliable. In social systems, where the consequences of model errors can be severe, having this knowledge is of even greater importance. Hence, the epistemic question of this work gets to the heart of various debates surrounding AI and its capabilities: How can we understand and measure the true capabilities of modern AI systems, which are so very impressive and yet lacking in fundamental ways at the same time [11]? What can we know about the quality of our models? Are our benchmarks suited to give insights into the intended tasks or do they project a false image of quality? How can we develop systems such that they work for everyone? Will naïve scaling solve all these problems or do we need to invest into entirely new approaches for evaluation within the scope of modern AI?

A prerequisite to answering RQ 1 positively is the validity of model validation: Without model validation we can not know wether a model is good or bad and without a valid model validation procedure we can not attain this knowledge. The almost exclusively used method for model validation in machine learning and AI is the ubiquitous train-test paradigm, i.e., the practice of estimating the generalization performance of a model on a test set distinct from the training set. Arguably, much of the breathtaking progress in machine learning has been driven by the success of this single experimental paradigm as it allows for the rapid validation and, therefore, improvement of models [10]. However, it is crucial to note that the train-test paradigm is inherently an inductive method that aims to *infer, not measure*, the generalization error of a model from its error on a test set. It is well known — dating back at least to Hume [28, 29] and formalized in the context of machine learning by Wolpert [68] — that it is not possible to justify the validity of such inductive inferences without further assumptions. This raises the question: is the train-test paradigm still valid for the combination of tasks and data sets considered in modern AI and under what assumptions is this the case?

Importantly, such assumptions should be minimal in terms of ontological commitments, i.e., meet *ontological parsimony* (or minimality), since (a) model validation results can not provide insights about validity in the real world if they are contingent on strong ontological assumptions (b) any assumptions that are required to ensure the validity of model validation can not be validated through the same method without circular reasoning. In traditional machine learning *settings*, these ontological commitments are placed entirely on the data collection process and, as such, the train-test paradigm is indeed suitable to *validate any model assumption* outside the data collection process. More concretely, under *active data collection*, i.e., when we actively control the data collection process, we can create large enough test sets that are (approximately) sampled i.i.d. from the target distribution. Under these conditions, it is well known that the train-test paradigm allows us to validate models simply via their performance on this test set — *without making any further ontological commitments*. This property is the beauty of the train-test paradigm and what makes it so valuable and successful.

However, domains in modern machine learning have become far too large to be covered via data sets in this active and controlled manner — the required effort would be prohibitively difficult and costly.

In lieu, *passive data collection* has become the predominant way to create data sets for modern AI systems. Here, data is collected without intervention from *some social system* that generates data within the domain of interest. For instance, rather than meticulously collecting independent samples from all possible facts in a domain, training and validation corpora for QA models are gathered from what has been published on the internet. Similarly, preferences of users are collected over items that a recommender system has pre-selected, rather than sampling them i.i.d. over all possible user-item pairs. Importantly, these sample generating systems need not correspond to the target data generating process, have their own internal dynamics, and are driven by complex interactions of their parts and social processes, e.g., well-known phenomena such as popularity bias [1], homophily [22, 37], or feedback loops [14].

Hence, I will ground RQ 1 in these conditions of current machine learning practice: *Under passive data collection from a social system, can model validation be valid or not?* To formalize the social systems with which an AI system interacts, I am taking a *complex systems* perspective and describe them as networks with well-established sampling biases and degree distributions. For these properties, I will show how they affect *necessary conditions* of test validity. These results can also be understood as a strengthening of the seminal *No Free Lunch* (NFL) theorems for supervised learning [68] in the context of social systems. While the NFL theorems show the impossibility of an assumption-free general purpose learning algorithm, a common criticism is that they need to assume an induction-hostile universe, i.e., full ontological neutrality [60]. In practice, where assuming a reasonably induction-friendly universe is common, the NFL theorems have had therefore limited impact. In contrast, the results of this work are grounded in current machine learning practice and considerably stronger: Even for non-trivial assumptions of an induction-friendly universe, model validation can be shown to be invalid when data is collected passively in social systems. In other words, there is *no free delivery service* of data for model validation in complex social systems. To discuss the above results, I will provide a synthesis of results from learning theory, social science, and complex systems — and combine them with new theoretical and empirical results on the validity of model validation. In particular, the *main contributions* of this paper are as follows:

> **Theorem 1** (Informal)**.** For passively collected data in complex social systems the train-test paradigm cannot be valid under ontological parsimony for the vast majority of the system. This includes widely considered variants of recommender systems and question answering.
>
> **Corollary 2** (Informal)**.** Naïve scaling and limited benchmarks are prohibitively inefficient to address theorem 1 and therefore not suited to attain test validity in these scenarios.
>
> **Supporting evidence.** Theoretical results are supported via experiments on the popular MOVIELENS benchmark where widely considered recommendation tasks are shown to be test-invalid.

The remainder of this paper proceeds as follows: Sections 2 and 3 formalize passive data collection in social systems and connect it to test validity. Section 4 develops theorem 1, corollary 2, and supporting evidence. Sections 5 and 6 discuss related work and implications for AI in social systems.

2 Passive data collection and inference tasks in social systems

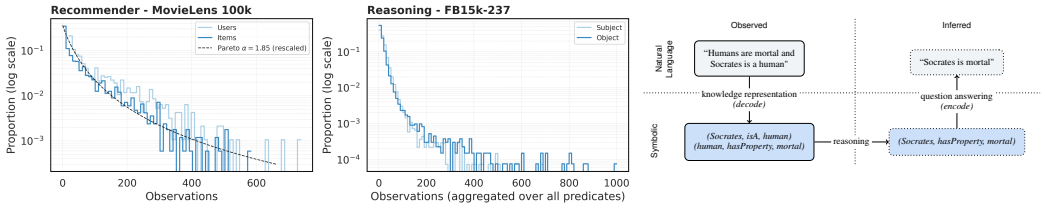
To construct validation data sets for large-scale domains, there exist currently two main practical approaches: (i) "scaling", i.e., indiscriminately collecting as much data as possible from some domain and (ii) manually constructing benchmarks of limited size that probe certain subareas of the domain. In the following, I will focus on formalizing (i) as passive data collection from social systems. Section 4 will then show that neither (i) nor (ii) can be solutions to the issues of this paper.

In sociology, a social system is often considered a pattern of networked interactions that exists between individuals, groups, or institutions [44]. For the purposes of this paper, I will consider a *social system* to be a pair (f, S) where $f : \mathcal{X} \to \mathcal{Y}$ is a *possible world of interactions* such that $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_n$ denotes the domain of interactions, $\mathcal{Y}$ denotes the set of outcomes (or labels) of an interaction, and $\mathsf{S} : \mathcal{X} \to [0, 1]$ denotes the *sampling distribution* of the system over interactions. Within this framework, *passive data collection* refers to sampling directly from S. This is in contrast to *active data collection* where we would aim to sample directly from the *target distribution* $\mathsf{T} : \mathcal{X} \to [0, 1]$ for an inference task, e.g., via simple random sampling, stratified sampling, etc.

In complex social systems, S is driven by social processes that lead to two characteristic properties of samples: (i) they are *biased* and (ii) they follow *heavy-tailed* or *power-law* distributions. The earliest

Table 1: **Inference settings based on passive data collection in complex social systems.**

	Domain $\mathcal{X}$	Possible world f	Sample distribution S	Target distribution T
Recommender systems	$\mathcal{U} \times \mathcal{I}$	User preferences	Probability of user interacting with item, heavy-tailed in $\mathcal{U}$ and $\mathcal{I}$	Uniform, $p_T(u,i) = 1/\lvert\mathcal{U} \times \mathcal{I}\rvert$
Symbolic reasoning	$\mathcal{S} \times \mathcal{P} \times \mathcal{O}$	Truth value of factoids	Probability of observing factoid, heavy-tailed in $\mathcal{S}$, $\mathcal{P}$, and $\mathcal{O}$	Uniform, $p_T(s,p,o) = 1/\lvert\mathcal{S} \times \mathcal{P} \times \mathcal{O}\rvert$



(a) Heavy-tailed samples

(b) Reasoning

Figure 2: (a) **Heavy-tailed samples** in recommender and reasoning datasets. (b) **Symbolic reasoning via LLMs.** To validate reasoning capabilities of LLMs, natural language has to be mapped to logical knowledge representations. This shows that validation of reasoning in LLMs is subject to the results of this paper. See also fig. 6b and supp. G.1.

work on (ii) is due to Simon [56], and has independently been discovered in multiple contexts. In fact, (ii) can often be understood as a consequence of (i), e.g., popularity bias leading to power-law distributions in social networks [5, 48]. See also supp. C for further discussion of these properties.

In the remainder, I will therefore focus on the presence of heavy-tailed distributions in $\mathcal{S}$ to understand how this ubiquitous property of social systems affects test validity. For this purpose, I will first introduce the concept of a sample graph, i.e., the observed interactions that we receive from S:

Definition 1 (Sample graph)**.** A data set $\mathcal{S} \sim \mathsf{S}^m \subset \mathcal{X}_1 \times \mathcal{X}_2$ of observed interactions induces a bipartite *sample graph* $G = (\mathcal{X}_1, \mathcal{X}_2, \mathcal{S})$ between entities of $\mathcal{X}_1$ and $\mathcal{X}_2$ where an edge indicates that the corresponding interaction has been observed. In the following, I will use $\mathcal{S}$ and G interchangeably.

For higher arity relations, definition 1 can easily be generalized to hypergraphs. For simplicity, I will focus on bipartite graphs in the following. In sample graphs, the heavy-tailed property of complex systems materializes then through their degree distribution. While the exact nature of these distributions is disputed [13], I will follow Voitalov et al. [64] and assume that node degrees in $\mathcal{S}$ follow a *regularly-varying power-law distribution*. Based on this observation, *passive data in complex social systems* will then refer to the following:

Definition 2 (Passive data in complex social systems)**.** Let $\mathcal{S} \sim \mathsf{S}^m$ be a sample graph drawn from sampling distribution S. Let K_1, K_2 denote random variables that model the degree distribution in $\mathcal{S}$ of nodes in $\mathcal{X}_1$ and $\mathcal{X}$, respectively. For passively collected data from complex social systems, I will then assume that K_1, K_2 follow regularly-varying power-law distributions, i.e.,

$$\mathbb{P}(K_1 > k) = u_1(k)k^{-\alpha_1} \quad \text{and} \quad \mathbb{P}(K_2 > k) = u_2(k)k^{-\alpha_2}$$

where $\alpha_i > 0$ are the tail indices and u_i are slowly varying functions such that $\lim_{x\to\infty} u(rx)/u(x) = 1$ for any $r > 0$. Higher arity relations are defined analogously. Next, I will show how passive data in social systems materializes in key inference settings (see also table 1).

Example 1 (Recommender Systems)**.** Recommender systems are concerned with inferring the true preferences of a user over all items from a set of revealed preferences sampled from S. As such they are a typical example for (f, S) where the target distribution T corresponds to the uniform distribution over all possible interactions. Importantly, S is typically influenced by social processes and sampling bias as well as heavy-tailed distributions are well documented in recommender systems. For instance, an important factor for sampling biases are feedback loops, e.g., that past recommendations influence which recommendations are shown in the future [34, 14]. Another source of sampling bias is user feedback, which is often biased towards items with high ratings [59], as well as popularity bias [1, 48]. Popularity bias leads directly to heavy-tailed distributions in the degree distribution of the sample graph [48, 5]. See also fig. 2a for evidence of this property on MovieLens.

Example 2 (Symbolic reasoning and QA)**.** Reasoning and question answering over symbolic knowledge representations are another key example for (f, S). In this setting, factoids are represented in form of *(subject, predicate, object)* triples and the task is to infer the truth value for *any* unknown factoid, i.e., for a uniform target distribution T. Importantly, while facts about the world itself do not need to be influenced by social processes, our available knowledge about them, i.e., S, often is. In addition to aspects such as popularity bias, causes for this can range from which questions are studied in science [35, 36], over how data is collected [31], to who has access to the internet and the ability to contribute to knowledge [67]. Consequently, heavy-tailed distributions are also well-documented in this setting. For instance, Steyvers et al. [61] showed that semantic networks typically follow heavy-tailed degree distributions. Similar distributions have been observed in large-scale knowledge graphs such as DBPEDIA [4], YAGO [62], FREEBASE [9], and WIKIDATA [65]. See also fig. 2a for evidence of this property on FB15K. Importantly, this setting applies to any reasoning task over factoids in general — irrespective of the data representation. For instance, the validation of reasoning capabilities for general purpose question answering in systems such as LLAMA [63, 21] and CHATGPT [47] needs to follow this blueprint. See also fig. 2b for an illustration.

3 Test validity

To answer RQ 1, I will focus on the test validity of *inference settings*, i.e., whether task, assumptions, and data allow for *any* valid validations at all. For this purpose, I will use a deductive approach: model validation is *valid* if it is a logical consequence of its assumptions that the difference between its estimate and the true generalization error is bounded with high probability. To formalize this, let $h, f : \mathcal{X} \to \mathcal{Y}$ denote functions that map from sample domain $\mathcal{X}$ to target domain $\mathcal{Y}$. For clarity, I will assume noise-free f and h. Furthermore, let $\mathcal{S} \sim \mathsf{S}^m = \{x_i\}_{i=1}^m$ denote a data set of m samples drawn from a sampling distribution $\mathsf{S} : \mathcal{X} \to [0, 1]$ and let $\mathcal{D} = \{(x, f(x)) : x \in \mathcal{S}\}$ denote its supervised extension. For notational convenience, I will also write $\mathcal{D} \sim \mathsf{S}^m$ when f is clear from context. In addition, let $\mathbb{A} \subseteq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$ be the set of all functions from $\mathcal{X}$ to $\mathcal{Y}$ that are consistent with some set of assumptions on f such as being low-rank. Next, note that $\mathbb{A}$ and $\mathcal{D}$ then induce a set of possible worlds as follows:

Definition 3 (Possible worlds)**.** Let $\mathbb{A}$ be a set of assumptions, $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$ a set of observations, and $f : \mathcal{X} \to \mathcal{Y}$. The set of *possible worlds* $\mathcal{F}$ is then the set of functions consistent with $\mathbb{A}$ and $\mathcal{D}$, i.e.,

$$\mathcal{F} = \{f \mid f \in \mathbb{A} \ \wedge \ \forall (x, y) \in \mathcal{D} : f(x) = y\}.$$

Furthermore, I will consider an inference setting $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ to be a set of assumptions $\mathbb{A}$, a *fixed* dataset $\mathcal{D} \sim \mathsf{S}^m$, a *target distribution* $\mathsf{T} : \mathcal{X} \to [0, 1]$ for which we want to make inferences, and an assumed *distribution over possible worlds* F. Note that if $\mathsf{S} \neq \mathsf{T}$, $\mathcal{D}$ can not be an i.i.d. sample from T. For further details and notation see supps. A and B.

Next, let X be a random variable over $\mathcal{X}$ and let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a positive loss function. The *risk* of hypothesis h with respect to a *single* world f is then denoted by

$$L_{fh}^{\mathsf{T}} = \mathbb{E}_{X \sim \mathsf{T}}[\ell(h(X), f(X))].$$

Furthermore, let θ denote *any* risk measure of a hypothesis h on some test set $\mathcal{T}$. For instance, θ could denote the empirical risk or a re-weighted estimator such as the Horvitz-Thompson adjusted empirical risk (see also table 4 in the supp. material). Hence, θ does not only cover the standard Monte-Carlo estimator for the i.i.d. setting, but also estimators used in counterfactual and causal settings. To determine the test-validity of an inference setting, I am then interested in bounding the difference between the estimated risk (θ) and the true risk of h (L_{fh}^{T}). Importantly, it is necessary to consider the risk of h relative to the distribution F over all possible worlds since no world $f \in \mathcal{F}$ can be excluded based on $\mathcal{D}$ and $\mathbb{A}$. Hence, test validity is defined as follows:

Definition 4 (Test validity)**.** Let $f \sim \mathsf{F}$ denote a distribution over possible worlds $\mathcal{F}$ and let $\mathcal{H}$ denote a hypothesis class. Furthermore, let L_{fh}^{T} denote the risk of hypothesis h for target distribution T and possible world f. Let $\theta \in \mathbb{R}_+$ denote any empirical risk measure of h on a test set. Then, $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ is (ϵ, δ)*-test-valid* (*test-invalid*) if θ's difference to L_{fh}^{T} can (cannot) be bounded accordingly, i.e.,

$$(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F}) \vDash \begin{cases} \exists \mathcal{H} \ \exists h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L_{fh}^{\mathsf{T}}| \leq \epsilon) \geq 1 - \delta & (\epsilon, \delta)\text{-test-validity} \\ \forall \mathcal{H} \ \forall h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L_{fh}^{\mathsf{T}}| > \epsilon) > \delta. & (\epsilon, \delta)\text{-test-invalidity} \end{cases}$$

The conditions in definition 4 for a valid validation setting are very mild since it requires only a single hypothesis class in which θ for a single hypothesis has bounded difference to the true risk with high probability. Since invalidity follows directly from validity via complement rule and negation, the conditions for a validation setting to be invalid are strong: For *any possible hypothesis class* it has to hold that the difference between θ and the true risk of *all hypotheses* can not be bounded with sufficient probability. Importantly, both are statements about an inference setting, i.e., the combination of assumptions, observed data, and target distribution, and not about a specific hypothesis (class). Furthermore, note that definition 4 implies realizability with regard to the assumptions: if $\{f \mid \forall (x, y) \in \mathcal{D} : f(x) = y\} \cap \mathbb{A} = \emptyset$, an inference setting is test-invalid since $\mathbb{P}(\emptyset) = 0$. However, definition 4 imposes no realizability or any other constraints on $\mathcal{H}$.

Next, note that definition 4 implies straightforward necessary conditions for test validity:

Corollary 1 (Necessary condition for test validity)**.** *Let* $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ *be an inference setting, let* $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ *be a positive loss function, and let* $\mathcal{H}$ *be a hypothesis class. Furthermore, let* $\theta \in \mathbb{R}_+$ *be any risk estimate for h. Then, if* $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ *is* (ϵ, δ)*-test-valid, it must hold that*

$$\exists \mathcal{H}\ \exists h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(L^{\mathsf{T}}_{fh} \leq \epsilon + \theta) \geq 1 - \delta.$$

Proof sketch. Corollary 1 follows simply via the monotonicity of probability, i.e., it holds that $1 - \delta \leq \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L^{\mathsf{T}}_{fh}| \leq \epsilon) \leq \mathbb{P}_{f \sim \mathsf{F}}(L^{\mathsf{T}}_{fh} \leq \epsilon + \theta)$. This holds for any risk measure θ, loss $\ell \in \mathbb{R}_+$ and hypothesis h. See supp. D for proof details. □

4 Test validity under passive data collection in complex systems

In the following, I will provide an overview of the main results as well as high-level proof sketches. For clarity, I will consider only binary relations $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ and possible worlds over unbounded output domains $f : \mathcal{X} \to \mathbb{R}$. For detailed proofs and discussion, as well as extensions to ternary relations and bounded domains, see supps. E to G. To meet ontological parsimony[1] and get insights into the validity of the train-test paradigm, I will focus on F being the uniform distribution U and $\mathbb{A}$ imposing only minimal assumptions on f.

Next, to derive bounds on the validity of inference settings in complex social systems, I will represent possible worlds f as partially observed matrices which are constructed as follows:

Definition 5 (Matrix representation)**.** For a function $f : \mathcal{X}_1 \times \mathcal{X}_2 \to \mathcal{Y}$ over *finite* sets of size $|\mathcal{X}_1| = n_1$ and $|\mathcal{X}_2| = n_2$, its *matrix representation* $\mathbf{F} \in \mathbb{R}^{n_1 \times n_2}$ is given via $\mathbf{F}_{ij} = f(x_i, x_j)$ for all $(x_i, x_j) \in \mathcal{X}_1 \times \mathcal{X}_2$.[2] In the following, I will use f and $\mathbf{F}$ interchangeably.

Using this matrix representation of a system, I will show in lemma 2 that the train-test paradigm is invalid if the rank of f, i.e., the complexity of the system, exceeds the k-connectivity of the sample graph $\mathcal{S}$ and if f is chosen uniformly from $\mathcal{F}$. Here, k-connectivity is defined as follows:

Definition 6 (k-core and k-connectivity)**.** The k-core (or core of order k) of a graph is its maximal subgraph such that all vertices are at least of degree k.[3] A graph is k-connected *if and only if* every vertex is in a core of order at least k.

Lemma 1 (Rank-k underdetermination)**.** *Let* $\mathbb{A} = \{f \mid rank(f) \leq k\}$*. Then, if* $\mathcal{S}$ *is not* k*-connected, the set of possible worlds* $\mathcal{F}$ *forms a non-empty vector space.*

Proof sketch. Since $\mathcal{S}$ is not k-connected, any f with $\text{rank}(f) = k$ can not be $\mathcal{S}$-isomeric. It then holds via [38, Lemma 5.1] that $\mathcal{F}$, i.e., the set of matrices of rank k or less that are consistent with $\mathcal{D}$, form a non-empty vector space. See supp. E for proof details. □

In the spirit of Occam's razor, higher ranks of f correspond to more complex possible worlds. Lemma 1 establishes then that if the k-connectivity of $\mathcal{S}$ does not match the complexity of the system f, the observations $\mathcal{S}$ do not constrain $\mathcal{F}$ sufficiently and a randomly chosen possible world can be arbitrarily different on the non-observed entries. Via corollary 1, lemma 1 implies then

[1] See also supp. B.2 for further discussion on the importance of ontological parsimony (minimality).

[2] This is trivially extended to higher arity functions using tensor representations. See also supp. G.1.

[3] Note that being in the k-core of $\mathcal{S}$ is a stronger condition than having degree k: A node can be outside the k-core even with a degree larger than k if enough of its neighbors are outside the k-core (see also fig. 6c)

that k-connectivity is necessary for test validity if ℓ belongs to the broad class of scalar Bregman divergences, i.e., widely used loss functions such as the square loss, the log loss, or the KL-divergence (see also table 5 in the supplementary material).

Lemma 2 (Rank-k test-invalidity)**.** *Let $\mathbb{A}$ be identical to lemma 1, let ℓ be a scalar Bregman divergence, let F be the uniform distribution over $\mathcal{F}$, and let T be the uniform distribution over $\mathcal{X}$. Furthermore, let $\theta \in \mathbb{R}_+$ be any risk estimator on a test set. Then, if $\mathcal{S}$ is not k-connected, $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ is test-invalid, i.e., it holds for any $\epsilon > 0$ that*

$$\forall \mathcal{H}\ \forall h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L^{\mathsf{T}}_{fh}| \leq \epsilon) = 0.$$

Proof sketch. If $\mathcal{S}$ is not k-connected, $\mathcal{F}$ is a vector space according to lemma 1. Lemma 2 follows then from corollary 1 for uniformly sampled $f \in \mathcal{F}$ and $h \in \mathcal{F}$ via a simple volume argument. For $h \notin \mathcal{F}$, the result follows again from $\mathcal{F}$ being a vector space via the generalized Pythagorean theorem for Bregman divergences [20, Eq. 2.3]. See supp. B.2 for proof details. □

The consequences of lemma 2 are non-trivial. Under ontological parsimony, it shows that passive data from complex social systems, i.e., the foundation of basically all large-scale AI tasks, can not be used to validate the quality of models if $\mathsf{S} \neq \mathsf{T}$. Clearly, no subset of $\mathcal{S}$, e.g., cross-validation, can fulfill this task either. Importantly, lemma 2 holds not only for empirical risk, but for *any* estimator on $\mathcal{D}$, including counterfactual estimators, i.e., methods which are exactly meant to address $\mathsf{S} \neq \mathsf{T}$. This illustrates that lemma 2 is not simply an out-of-distribution or counterfactual estimation problem. Rather, it is caused by a combination of out-of-distribution ($\mathsf{S} \neq \mathsf{T}$) and insufficient data (k-connectivity $<$ rank(f)). Next, I will connect these results to the main result of this work.

Theorem 1 (Test validity in complex social systems)**.** *Let $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ be identical to lemma 2. Furthermore, let $\mathcal{S} \sim \mathsf{S}^m$ where S follows power-law distributions such that the degrees of $x \in \mathcal{X}_i$ in the sample graph $\mathcal{S}$ are drawn i.i.d. from a regularly-varying power-law distribution $\mathbb{P}(\deg(x) > k) = u(k)k^{-\alpha_i}$. Furthermore, let $n_i = |\mathcal{X}_i|$ be the size of domain $\mathcal{X}_i$. Then, the number V_i of nodes in $\mathcal{X}_i$ for which test validity holds decreases with a power-law decay in rank$(f) = k$, i.e,*

$$\mathbb{E}[V_i] \leq n_i u(k) k^{-\alpha_i}.$$

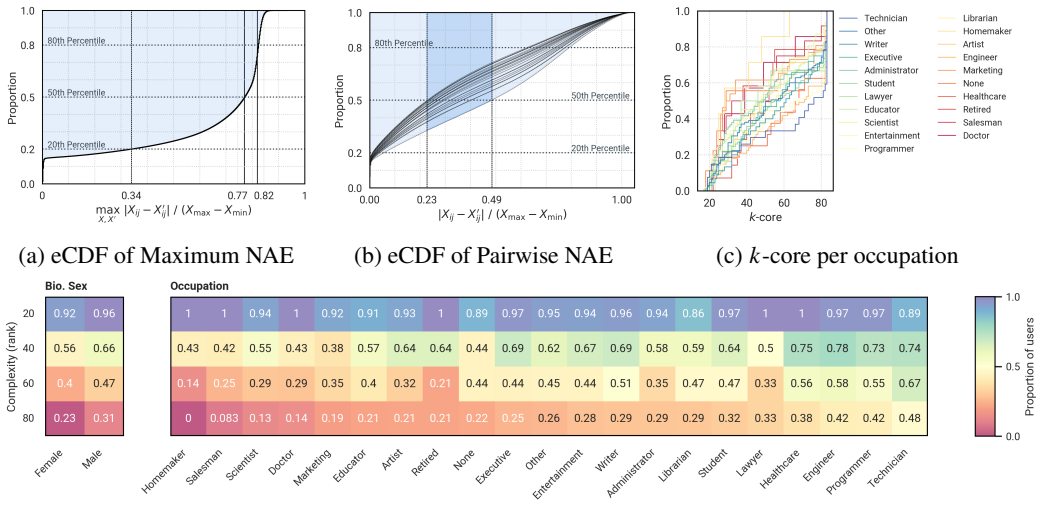
Proof sketch. Test validity requires the k-connectivity of $\mathcal{S}$ to be greater or equal to rank(f) via lemmas 1 and 2. Hence, only subgraphs where all vertices are at least of degree k can be valid. Theorem 1 follows then via the expected number of nodes with degree at least k in $\mathcal{X}_i$, i.e, $\mathbb{E}[V_i] = \sum_{x \in \mathcal{X}_i} \mathbb{P}(\deg(x) \geq k)$. □

For heavy-tailed distributions, most nodes will be outside the required k-core for even moderately complex worlds. Hence, theorem 1 shows that the train-test paradigm cannot be valid under ontological parsimony for the vast majority of nodes in realistic social systems. Table 2 illustrates this using parameters that match the well-known Book Crossing dataset.

An immediate next question is then if the issues raised by theorem 1 can simply be solved by scaling, i.e., by collecting more data from S — or via manually constructed benchmarks such as BigBench [58] to extrapolate from their results to the risk on T. Corollary 2 answers both questions via lemma 1 (see supp. H for a detailed discussion and proof): (i) For *scaling*, we can ask how many draws from S would be necessary such that all nodes are within the k-core of $\mathcal{S}$ with high probability, i.e., how many samples are needed until arriving at a valid test setting. While there exists no easily computable solution to this problem, we can compute a (weak) lower bound by asking how many samples from S are needed to sample a random node in $\mathcal{X}_i$ *once*. (ii) For *benchmarks*, we can ask how many nodes would need *at least one* additional data point to arrive at a valid test setting, i.e., how much manual data collection is at least needed to create a benchmark that extrapolates to T.

Table 2: **Inefficiency of scaling and benchmarks; validity coverage for the Pareto distribution.**

α	$x_{\min}$	$\lvert\mathcal{X}\rvert$	**Scaling** Samples needed to increase k-core of random node	**Benchmarks** Nodes with less than 100 observations
2.5	5	10^7	$\mathbb{E}_{i \sim \mathsf{U}}[T_i] \geq (\lvert\mathcal{X}\rvert/2)^{\alpha+1}/(\alpha x_{\min}^{\alpha}) = 2 \cdot 10^{21}$	$\mathbb{E}[N] = \lvert\mathcal{X}\rvert(1 - (x_{\min}/x)^{\alpha}) > 9.9 \cdot 10^6$
α	$x_{\min}$	$\lvert\mathcal{X}\rvert$	**Book Crossing** [69] Fraction of users with large enough degrees such that train-test measures and inferences are valid	
2.38	8	10^5	Rank 8: 100%, Rank 10: 58.8%, Rank 20: 11.3%, Rank 100: 0.2%	



(a) eCDF of Maximum NAE (b) eCDF of Pairwise NAE (c) k-core per occupation

(d) Test-validity per demographic group and model complexity

Figure 3: **MovieLens 100k experiments** (a) Empirical CDF (eCDF) of maximum NAE over possible worlds. Area over the curve (expected error) shaded. (b) eCDF of NAE for pairs of possible worlds. (c) eCDF k-coreper demographic group. (d) Proportion of users for which test-validity holds relative to the rank of f.

Corollary 2 (Inefficiency of scaling and benchmarks)**.** *Let* $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ *and* S *be identical to theorem 1. Furthermore, let (i)* T_i *denote the expected number of samples from* S *until node* $x_i \in \mathcal{X}$ *is sampled, and let (ii)* N_j *denote the number of nodes in* $\mathcal{X}_j$ *with less then* k *samples. Then,* T_i *scales at least* polynomially *and* N_j *scales lineraly in* the size of the domain $|\mathcal{X}|$. *Specifically,*

$$\mathbb{E}_{i\sim\mathsf{U}\{1,|\mathcal{X}|\}}[T_i] \geq (|\mathcal{X}|/2)^{\alpha+1}/(\alpha x_{min}^{\alpha}), \quad \textit{and} \quad \mathbb{E}[N_j] = |\mathcal{X}|(1-(x_{min}/x)^{\alpha})$$

Clearly, sampling from S is highly inefficient to overcome the issues raised by theorem 1 since (i) it is extremely difficult to get successful samples from the heavy tail (rare events) and (ii) covering all nodes outside sufficiently large k-cores in selective benchmarks is prohibitively expensive. See also table 2 for examples of these aspects for typical distributions in complex social systems.

Experimental evidence To illustrate the real consequences of the previous theoretical results, I will now provide experimental evidence based on the MovieLens 100k dataset [27], a critical benchmark that has, for years, been widely-used in recommender systems research. As predicted by lemma 2, I will show that there exist possible worlds of low complexity that all explain the observed data equally well but are widely different on the unobserved data. Hence, *any* quality metric that is inferred on this benchmark, or subsets of it, can not be informative about the true generalization error. For this purpose, I fit $p = 100$ matrices of rank $k = 50$ to the observed data $\mathcal{D}$. All matrices, or possible worlds, fit the observed data and rank constraint with error below 10^{-3} and 10^{-2}, respectively. See supp. I.1 for details. For a pair of possible worlds (f, f') , I compute then the normalized absolute error (NAE) for each *unobserved* entry $(i, j) \notin \mathcal{S}$ via $\text{NAE}(f_{ij}, f'_{ij}) = |f_{ij} - f'_{ij}|/(f_{\max} - f_{\min})$. This informs us about how different pairs of possible worlds can be on the unobserved data. Figures 3a and 3b shows the empirical CDF (eCDF) of the NAE over unobserved entries for such pairwise comparisons of possible worlds as well as the worst-case over all worlds per entry. From fig. 3a, it can be seen that the worst case error across possible worlds per entry is substantial for the vast majority of unobserved entries. For instance, for 50% of entries the NAE is above 77% of the worst case error. For arbitrary pairs of possible worlds, the situation is similar, where, depending on the particular pair of worlds, the NAE is between 23% to 49% for 50% of entries. Furthermore, the area over the eCDF curves in fig. 3b corresponds directly to the risk for a pair of possible worlds and is again substantial for all pairs (see supp. I.2 for details). Since any possible world can be the "true" world this shows again that the test error for any subset of this benchmark can not be informative for the true generalization error of this task.

In addition to the NAE, fig. 3c shows the cumulative distribution of users within cores of order k per demographic group for MovieLens 100k. It can be seen that the cumulative distribution can vary significantly between different demographics. For instance, while only 25% of "homemakers" are in

a k-core larger than 50, 40% of "technicians" are in a k-core larger than 80. It follows from lemma 2, that test-validity will therefore also vary significantly between demographic groups (if we assume that there are no significant differences in the complexity of preferences between groups). Figure 3d illustrates this point by showing the proportion of users for which test-validity holds relative to the rank of a model. It can be seen that there exist clear differences already for moderately complex worlds. For instance, for a model of rank 60, test-validity would hold for 67% of "technicians" while it would only hold for 14% of "homemakers". Clearly, this has important implications for fairness, bias, and whether recommender systems *work for everyone*.

5 Related work

The no-free-lunch theorems for machine learning [68, 60] share important similarities to this work as both consider the expected risk over possible worlds. However, the results in this paper are stronger and directly applicable to current machine learning practice. While the NFL theorems consider the performance over all possible worlds without any restrictions — an assumption that is too restrictive in most instances — the results of this paper show that even for relatively strong assumptions about the set of possible worlds, e.g., low-rank structures, valid model validation is not generally possible for passive data collection in complex social systems. In motivation, this paper is also related to the works [19, 7, 25, 54, 41, 66] which study outcomes of underspecification in ML pipelines, model multiplicity and Rashomon sets. In the restricted context of personalized prediction, Monteiro Paes et al. [46], discusses related limits to testing and estimation. Schaeffer et al. [52] discuss whether seemingly emergent capabilities of LLMs are rather a result of insufficient metrics. In statistics, Meng [43] analyzed a scaling-related question similar to this paper: Given a carefully collected survey with low response rate (small data) or a large, self-reported dataset without data curation (big data), which dataset should one trust more to estimate population averages? Outside machine learning, validity theory has a long history in fields such as psychology and sociology. Here, test validity is considered a measure of the degree to which a test measures what it is intended to measure [18] and has been studied extensively in the context of psychological tests [45] and educational testing [32]. Increasingly, these notions of validity, have also been considered in machine learning [17, 51, 50, 2].

With regard to technical tools, this paper is also closely related to prior work in matrix completion. For instance, [33] studied the problem of unique and finite completability of matrices and derived similar k-core related bounds using determinantal varieties and algebraic geometry. Srebro et al. [57] studied the problem of matrix completion based on non-uniform samples such as power-laws but assume that $\mathsf{S} = \mathsf{T}$. Meka et al. [42] focused on power-law samples for $\mathsf{S} \neq \mathsf{T}$ and, consistent with this work, require at least k samples per row and column to guarantee completability of a rank-k matrix. Cheng et al. [15] derive similar results based on graph k-connectivity. Related to non-i.i.d. observations, [38] developed a framework to provide necessary conditions for matrix completion under deterministic sampling. Lemma 1 is based on these results. Different to these prior works, I provide formal impossibility results for test validity based on passive data in complex social systems. This allows to gain rigorous insights into the epistemic limits of what we can know based on this form of data collection. See also supp. J for further related work.

6 Discussion

The results in this paper provide new insights into the validity of the train-test paradigm when data is passively collected from complex social systems. In particular, I have shown that there exists *no free delivery service* of data that allows for test validity on a global scale in this setting. While valid inferences are possible with respect to the sampling distribution S and within high k-cores, they are unlikely if T extends to the entirety of the system. Hence, test validity depends on the interplay between task (T), the complexity of the system ($\mathbb{A}$), and the k-connectivity of the sample graph ($\mathcal{S}$) underlying the observed data ($\mathcal{D}$), what is a *combinatorial* property of the data. These results are attained by establishing novel *necessary conditions* for which validation is possible. As AI systems are increasingly applied in conditions for which sufficient conditions of validity are difficult to guarantee, understanding such minimal conditions can provide guidelines into developing better and more robust systems. Importantly, it can help to demarcate inference goals that are not meaningful from ones that are attainable. It helps to understand the limits of what we can know and

which questions are futile to ask. This work provides a first step in this direction by establishing such epistemic limits of AI in complex social systems.

Furthermore, I have shown that the sub-system for which valid inferences are possible shrinks rapidly with the complexity of the system and that a naïve application of the scaling paradigm is prohibitively inefficient to overcome these validity issues. As a consequence, solving many complex AI tasks are unlikely to come for free through scaling or for cheap through extrapolating from limited small-scale benchmarks. Instead, there exists an inherent trade-off between data quality, quantity, and task complexity. If we want to avoid asking AI systems to solve simpler tasks (e.g., non-out-of-distribution or smaller scope), new data curation efforts are likely needed. Due to the substantial amount of data that would have to be collected, centralized data collection is often infeasible to overcome the validity issues of this paper. Instead, decentralized methods such as *participatory data curation* could provide a way forward. This aligns with insights from fairness which also highlight the need for participatory methods in data collection [31]. Similar arguments apply to the importance of open science and open-source models in this context.

Importantly, the theoretical results of this paper also provide direct insights into how to improve data collection for model validation via its k-core conditions. In particular, lemma 1 and corollary 2 imply two clear objectives for targeted data collection: (a) collecting data points that increase the k-connectivity of the sample graph and (b) collecting data points that increase the size of the $\text{rank}(f)$-core of the sample graph, where $\text{rank}(f)$ is the complexity of the world that we want to assume. Pursuing (a) would increase the complexity of the world that can be assumed such that model validation is still valid for the entire sample graph, while pursuing (b) would increase the size of the subgraph for which a $\text{rank}(f) = k$ assumption would still yield valid model validation. Hence, both objectives are based on the k-core conditions of this work and can be computed from a given sample graph. Creating new mechanisms for efficient data collection based on these insights is therefore a very promising avenue for future work.

Acknowledgments

I gratefully acknowledge the valuable feedback from Léon Bottou, Smitha Milli, Tina Eliassi-Rad, Mark Tygert, and anonymous reviewers which all helped to improve various versions of this paper.

References

[1] Himan Abdollahpouri. "Popularity bias in ranking and recommendation". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 529–530.

[2] Rediet Abebe. *Algorithms on the Bench: Examining Validity of ML Systems in the Public Sphere*. 2022.

[3] Chris Anderson. *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. en. June 23, 2008. (Visited on 02/27/2024).

[4] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. "DBpedia: A Nucleus for a Web of Open Data". In: *The Semantic Web*. Springer Berlin Heidelberg, 2007, pp. 722–735.

[5] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *Science* 286 (5439 1999), pp. 509–512.

[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. en. MIT Press, Dec. 19, 2023. 340 pp.

[7] Emily Black, Manish Raghavan, and Solon Barocas. "Model Multiplicity: Opportunities, Concerns, and Solutions". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea). FAccT '22. New York, NY, USA: Association for Computing Machinery, June 20, 2022, pp. 850–863. (Visited on 02/27/2024).

[8] Mathieu Blondel, Quentin Berthet, Marco Cuturi, Roy Frostig, Stephan Hoyer, Felipe Llinares-López, Fabian Pedregosa, and Jean-Philippe Vert. "Efficient and Modular Implicit Differentiation". In: *arXiv preprint arXiv:2105. 15183* (2021).

[9] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. "Freebase: a collaboratively created graph database for structuring human knowledge". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data* (Vancouver, Canada). SIGMOD '08. New York, NY, USA: Association for Computing Machinery, June 9, 2008, pp. 1247–1250. (Visited on 12/31/2023).

[10] Léon Bottou. *Two big challenges in machine learning*. 2015.

[11] Léon Bottou and Bernhard Schölkopf. "Borges and AI". In: *arXiv [cs.CL]* (Sept. 27, 2023).

[12] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. *JAX: composable transformations of Python+NumPy programs*. Comp. software. Version 0.3.13. 2018.

[13] Anna D Broido and Aaron Clauset. "Scale-free networks are rare". In: *Nat. Commun.* 10 (1 2019), pp. 1–10.

[14] Allison J B Chaney, Brandon M Stewart, and Barbara E Engelhardt. "How algorithmic confounding in recommendation systems increases homogeneity and decreases utility". In: *Proceedings of the 12th ACM conference on recommender systems*. 2018, pp. 224–232.

[15] Dehua Cheng, Natali Ruchansky, and Yan Liu. "Matrix completability analysis via graph k-connectivity". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, 2018, pp. 395–403.

[16] Lu Cheng, Kush R Varshney, and Huan Liu. "Socially Responsible AI Algorithms: Issues, Purposes, and Challenges". en. In: *jair* 71 (Aug. 28, 2021), pp. 1137–1181. (Visited on 12/25/2023).

[17] Amanda Coston, Anna Kawakami, Haiyi Zhu, Ken Holstein, and Hoda Heidari. "A validity perspective on evaluating the justified use of data-driven decision-making algorithms". In: *arXiv [cs.LG]* (June 29, 2022). (Visited on 11/01/2024).

[18] Lee J Cronbach and Paul E Meehl. "Construct validity in psychological tests". In: *Psychol. Bull.* 52 (4 1955), p. 281.

[19] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D Sculley. "Underspecification presents challenges for credibility in modern machine learning". In: *J. Mach. Learn. Res.* 23 (1 Jan. 1, 2022), pp. 10237–10297.

[20] Inderjit S Dhillon and Joel A Tropp. "Matrix nearness problems with Bregman divergences". In: *SIAM J. Matrix Anal. Appl.* 29 (4 Jan. 2008), pp. 1120–1146.

[21] Abhimanyu Dubey et al. "The Llama 3 herd of models". In: *arXiv [cs.AI]* (July 31, 2024). (Visited on 10/30/2024).

[22] Francesco Fabbri, Francesco Bonchi, Ludovico Boratto, and Carlos Castillo. "The Effect of Homophily on Disparate Visibility of Minorities in People Recommender Systems". en. In: *ICWSM* 14 (May 26, 2020), pp. 165–175. (Visited on 12/25/2023).

[23] Vitaly Feldman. "Does learning require memorization? a short tale about a long tail". en. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. STOC '20: 52nd Annual ACM SIGACT Symposium on Theory of Computing (Chicago IL USA). New York, NY, USA: ACM, June 22, 2020. (Visited on 04/28/2024).

[24] Vitaly Feldman and Chiyuan Zhang. "What Neural Networks Memorize and Why: Discovering the Long Tail via Influence Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by H Larochelle, M Ranzato, R Hadsell, M F Balcan, and H Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 2881–2891.

[25] Aaron Fisher, C Rudin, and F Dominici. "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously". In: *J. Mach. Learn. Res.* 20 (177 Jan. 4, 2018), pp. 1–81. (Visited on 04/28/2024).

[26] Philippe Flajolet, Danièle Gardy, and Loÿs Thimonier. "Birthday paradox, coupon collectors, caching algorithms and self-organizing search". en. In: *Discrete Appl. Math.* 39 (3 Nov. 11, 1992), pp. 207–229.

[27] F Maxwell Harper and Joseph A Konstan. "The MovieLens Datasets: History and Context". In: *ACM Trans. Interact. Intell. Syst.* 5 (4 Dec. 22, 2015), pp. 1–19.

[28] David Hume. *A Treatise of Human Nature (I) of the understanding*. Vol. 1. 1739. (Visited on 12/21/2023).

[29] David Hume. *An Enquiry Concerning Human Understanding*. 1748.

[30] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto, and Pascale Fung. "Survey of Hallucination in Natural Language Generation". In: *arXiv [cs.CL]* (Feb. 8, 2022).

[31] Eun Seo Jo and Timnit Gebru. "Lessons from Archives: Strategies for Collecting Sociocultural Data in Machine Learning". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain). FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, pp. 306–316.

[32] Michael T Kane. "Validating the interpretations and uses of test scores". en. In: *J. Educ. Meas.* 50 (1 Mar. 2013), pp. 1–73.

[33] Franz J Király, Louis Theran, and Ryota Tomioka. "The algebraic combinatorial approach for low-rank matrix completion". In: *J. Mach. Learn. Res.* 16 (1 2015), pp. 1391–1436.

[34] Karl Krauth, Yixin Wang, and Michael I Jordan. "Breaking Feedback Loops in Recommender Systems with Causal Inference". In: *arXiv preprint arXiv:2207. 01616* (2022).

[35] Thomas S Kuhn. *The structure of scientific revolutions*. en. 2nd Edition. Chicago, IL: University of Chicago Press, Apr. 1, 1970. 210 pp.

[36] Hugh Lacey. *Is Science Value Free?: Values and Scientific Understanding*. Taylor Francis, 2005. (Visited on 01/24/2024).

[37] David Liu, Virginie Do, Nicolas Usunier, and Maximilian Nickel. "Group fairness without demographics using social networks". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM Conference on Fairness, Accountability, and Transparency (Chicago, IL, USA, 2023). FAccT '23. New York, NY, USA: Association for Computing Machinery, June 12, 2023, pp. 1432–1449. (Visited on 04/28/2024).

[38] Guangcan Liu, Qingshan Liu, Xiao-Tong Yuan, and Meng Wang. "Matrix completion with deterministic sampling: Theories and methods". In: *IEEE Trans. Pattern Anal. Mach. Intell.* 43 (2 2019), pp. 549–566.

[39] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. "Towards Out-Of-Distribution Generalization: A Survey". In: *arXiv [cs.LG]* (Aug. 31, 2021).

[40] Benjamin M Marlin and Richard S Zemel. "Collaborative prediction and ranking with non-random missing data". In: *Proceedings of the third ACM conference on Recommender systems*. 2009, pp. 5–12.

[41] Charles T Marx, Flavio du Pin Calmon, and Berk Ustun. "Predictive Multiplicity in Classification". In: *arXiv [cs.LG]* (Sept. 14, 2019).

[42] Raghu Meka, Prateek Jain, and Inderjit Dhillon. "Matrix Completion from Power-Law Distributed Samples". In: *Advances in Neural Information Processing Systems*. Ed. by Y Bengio, D Schuurmans, J Lafferty, C Williams, and A Culotta. Vol. 22. Curran Associates, Inc., 2009.

[43] Xiao-Li Meng. "Statistical Paradises And Paradoxes In Big Data (I) Law Of Large Populations, Big Data Paradox, And The 2016 US Presidential Election". In: *Ann. Appl. Stat.* 12 (2 2018), pp. 685–726.

[44] Merriam Webster Dictionary. *Social system*. en. (Visited on 10/30/2024).

[45] Samuel Messick. "Meaning and Values in Test Validation: The Science and Ethics of Assessment". In: *Educ. Res.* 18 (2 1989), pp. 5–11.

[46] Lucas Monteiro Paes, Carol Long, Berk Ustun, and Flavio Calmon. "On the Epistemic Limits of Personalized Prediction". In: *Advances in Neural Information Processing Systems*. Ed. by S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh. Vol. 35. Curran Associates, Inc., 2022, pp. 1979–1991.

[47] OpenAI et al. "GPT-4 Technical Report". In: *arXiv [cs.CL]* (Mar. 15, 2023).

[48] Fragkiskos Papadopoulos, Maksim Kitsak, M Ángeles Serrano, Marián Boguná, and Dmitri Krioukov. "Popularity versus similarity in growing networks". In: *Nature* 489 (7417 2012), p. 537.

[49] Daniel L Pimentel-Alarcón, Nigel Boston, and Robert D Nowak. "A characterization of deterministic sampling patterns for low-rank matrix completion". In: *IEEE J. Sel. Top. Signal Process.* 10 (4 2016), pp. 623–636.

[50] Deborah Raji. *There's more to data than distributions*. Mar. 31, 2022. (Visited on 03/31/2022).

[51] Benjamin Recht. *Machine Learning has a validity problem*. 2022. (Visited on 03/15/2022).

[52] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. "Are Emergent Abilities of Large Language Models a Mirage?" In: *Thirty-seventh Conference on Neural Information Processing Systems*. Nov. 2, 2023. (Visited on 05/13/2024).

[53] Tobias Schnabel, Adith Swaminathan, Ashudeep Singh, Navin Chandak, and Thorsten Joachims. "Recommendations as Treatments: Debiasing Learning and Evaluation". In: *Proceedings of The 33rd International Conference on Machine Learning*. Ed. by Maria Florina Balcan and Kilian Q Weinberger. Vol. 48. Proceedings of Machine Learning Research. New York, New York, USA: PMLR, 2016, pp. 1670–1679.

[54] Lesia Semenova, Cynthia Rudin, and Ronald Parr. "On the existence of simpler machine learning models". en. In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency (Seoul Republic of Korea). New York, NY, USA: ACM, June 21, 2022. (Visited on 04/28/2024).

[55] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. en. Cambridge University Press, May 19, 2014. 415 pp.

[56] Herbert A Simon. "On a class of skew distribution functions". In: *Biometrika* 42 (3/4 1955), pp. 425–440.

[57] Nathan Srebro and Russ R Salakhutdinov. "Collaborative Filtering in a Non-Uniform World: Learning with the Weighted Trace Norm". In: *Advances in Neural Information Processing Systems*. Ed. by J Lafferty, C Williams, J Shawe-Taylor, R Zemel, and A Culotta. Vol. 23. Curran Associates, Inc., 2010.

[58] Aarohi Srivastava et al. "Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models". In: *Transactions on Machine Learning Research* (Jan. 19, 2023). (Visited on 01/06/2024).

[59] Harald Steck. “Training and testing of recommender systems on data missing not at random”. In: *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. 2010, pp. 713–722.

[60] Tom F Sterkenburg and Peter D Grünwald. “The no-free-lunch theorems of supervised learning”. In: *Synthese* 199 (3-4 2021), pp. 9979–10015.

[61] Mark Steyvers and Joshua B Tenenbaum. “The Large-scale structure of semantic networks: Statistical analyses and a model of semantic growth”. In: *Cogn. Sci.* 29 (1 2005), pp. 41–78.

[62] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. “Yago: a core of semantic knowledge”. In: *Proceedings of the 16th international conference on World Wide Web* (Banff, Alberta, Canada). WWW ’07. New York, NY, USA: Association for Computing Machinery, May 8, 2007, pp. 697–706. (Visited on 12/31/2023).

[63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. “Llama 2: Open Foundation and Fine-Tuned Chat Models”. In: *arXiv [cs.CL]* (July 18, 2023).

[64] Ivan Voitalov, Pim van der Hoorn, Remco van der Hofstad, and Dmitri Krioukov. “Scale-free networks well done”. In: *Phys. Rev. Res.* 1 (Nov. 5, 2018), p. 033034.

[65] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Commun. ACM* 57 (10 Sept. 23, 2014), pp. 78–85.

[66] Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. “Predictive multiplicity in probabilistic classification”. en. In: *Proc. Conf. AAAI Artif. Intell.* 37 (9 June 26, 2023), pp. 10306–10314. (Visited on 04/28/2024).

[67] Wikipedia contributors. *Wikipedia:Who writes Wikipedia*. (Visited on 01/24/2024).

[68] David H Wolpert. “The lack of a priori distinctions between learning algorithms”. In: *Neural Comput.* 8 (7 1996), pp. 1341–1390.

[69] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. “Improving recommendation lists through topic diversification”. en. In: *Proceedings of the 14th international conference on World Wide Web - WWW ’05*. the 14th international conference (Chiba, Japan). New York, New York, USA: ACM Press, 2005. (Visited on 01/04/2024).

Supplementary Information

A Notation

Random variables are denoted by italic uppercase letters, e.g., L, S, X. Sets are denoted by calligraphic uppercase letters, e.g., $\mathcal{X}, \mathcal{S}$. Constants are indicated with lowercase greek letters, e.g., ϵ, ρ. Functions and scalar are denoted by lowercase letters, e.g., f, g, h and x, y. Matrices and higher-order tensors are indicated with bold uppercase letters, e.g., $\mathbf{F}, \mathbf{U}$.

Table 3: Notation

Concept	Notation
Possible world	$f : \mathcal{X} \to \mathcal{Y}$
Hypothesis	$h : \mathcal{X} \to \mathcal{Y}$
Loss	$\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$
Sample distribution	$\mathsf{S} : \mathcal{X} \to [0, 1]$
Target distribution	$\mathsf{T} : \mathcal{X} \to [0, 1]$
Social system	(f, S)
Sample graph	$\mathcal{S} \sim \mathsf{S}^m$
Test set	$\mathcal{D} = \{(x, f(x)) : x \in \mathcal{S}\}$
Risk	$L^{\mathsf{T}}_{fh} = \mathbb{E}_{X \sim \mathsf{T}}[\ell(h(X), f(X))]$
Estimated risk	$\theta \in \mathbb{R}_+$

B Validity framework

In this work, I am interested in the validity of *inference settings*, i.e., whether assumptions and observations allow for *any* valid inferences at all. To formalize this, I will take the following high-level approach:

Inference setting An inference setting consists of a set of assumptions $\mathbb{A}$, a *fixed* dataset $\mathcal{D}$ which is collected from a sampling distribution S, and a target distribution T for which we want to make inferences. Note that S is not guaranteed to be identical to T. Hence, we're concerned with out-of-distribution generalization settings.

Expected risk over possible worlds Assumptions $\mathbb{A}$ and observed data $\mathcal{D}$ define a set of possible worlds $\mathcal{F}$ that is consistent with $\mathbb{A}$ and $\mathcal{D}$. Given a probability distribution F over $\mathcal{F}$, I am then interested in the expected risk over all possible worlds that are consistent with $\mathbb{A}$ and $\mathcal{D}$.

Validity An inference setting is valid, if the expected risk over possible worlds can be bounded meaningfully at all, i.e., if there exists *at least one* hypothesis class for which the generalization error of at least *a single* hypothesis can be bounded sufficiently.

To approach the question of validity, learning theory has traditionally focused nearly exclusively on *sufficient conditions* for valid inferences. Under active data collection, i.e., in scenarios where one can control exactly how data is collected, sufficient conditions are highly attractive since they provide exact specifications for inferences to be valid with high probability. However, under passive data collection, the situation is reversed. Sufficient conditions for the validity of inferences usually place highly restrictive demands on the data collection process (e.g., i.i.d. samples or simple random sampling) which are challenging to satisfy even when data is collected carefully in an active way. Since passive data collection, by definition, exerts no control over the sample generating process, these sufficient conditions are not met with near certainty. For this reason, I am focusing here on *necessary conditions* for validity, i.e., conditions that must always be satisfied for inferences to be valid. Under passive data collection, necessary conditions can provide important insights since they need to hold for any data collection process or, conversely, can be used to identify scenarios where inferences are not valid with high probability.

B.1 Connection to No-Free-Lunch theorems

The validity framework of section 3 and the No-Free-Lunch theorems are closely connected. First, consider the *expected risk over all possible worlds* relative to F, i.e.,

$$\mathbb{E}_{f\sim\mathsf{F}}\left[L^{\mathsf{T}}_{fh}\right] = \mathbb{E}_{f\sim\mathsf{F}}\,\mathbb{E}_{X\sim\mathsf{T}}[\ell(h(X), f(X))]. \tag{1}$$

Equation (1) is then akin to the objectives considered in the seminal *No Free Lunch* (NFL) theorems [68, 60]. For instance, the NFL theorem for supervised learning can be written as $\forall\Lambda : \mathbb{E}_{f\sim\mathsf{U}}\,\mathbb{E}_{X\sim\mathsf{T}}[\ell(h_{\Lambda(\mathcal{D})}(X), f(X))] = 1/2$, where U is the uniform distribution over all possible worlds in an assumption-free setting (i.e., $\mathbb{A} = \emptyset$), ℓ is the $0/1$-loss, and $h_{\Lambda(\mathcal{D})}$ is the hypothesis derived from a finite sample $\mathcal{D}$ with algorithm Λ. In contrast to the NFL theorems — where $\mathbb{A} = \emptyset$ implies an induction-hostile universe — my focus is on induction-friendly settings ($\mathbb{A} \neq \emptyset$) but where $\mathcal{D}$ is sampled from a complex social system. Since L^{T}_{fh} is a non-negative random variable, we can then connect definition 4 and eq. (1) via upper and lower bounds based on Markov's inequality.

Definition 7 (Markov's inequality)**.** Let X be a non-negative random variable and $a > 0$. Then

$$\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a.$$

Hence, it follows that the expected risk over all possible worlds is large for invalid settings since it holds that

$$\mathbb{E}_{f\sim\mathsf{F}}[L^{\mathsf{T}}_{fh}] \geq \epsilon \cdot \mathbb{P}_{f\,\sim\,\mathsf{F}}(L^{\mathsf{T}}_{fh} > \epsilon)$$

B.2 Importance of ontological parsimony and test validity in the i.i.d. setting

The strong appeal of the train-test paradigm is that, with careful data collection, we require no further ontological assumptions to ensure the validity of the model validation procedure. In particular, if we have a test set that is sampled independently from T, it follows straightforwardly from Hoeffding's inequality that we can meaningfully bound the approximation error over this test set [55, Theorem 11.1]. Let $\mathcal{T} \sim \mathsf{T}^m$ be a test set of size m, sampled i.i.d. from the target distribution T. Then, it holds that

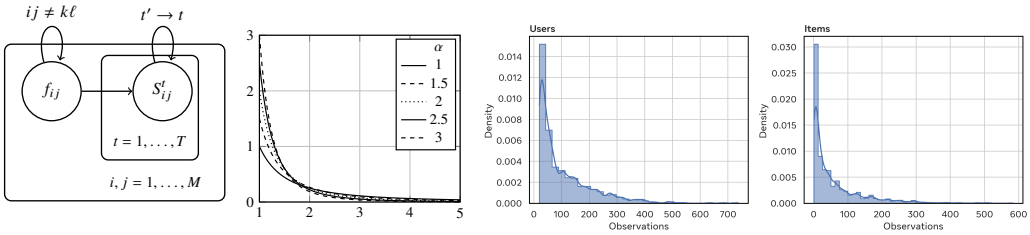
$$\mathbb{P}_{\mathcal{T}\sim\mathsf{T}^m}\left(\left|L^{\mathcal{T}}_{hf} - L^{\mathsf{T}}_{hf}\right| \leq \sqrt{\frac{\log(2/\delta)}{2m}}\right) \geq 1 - \delta.$$

Importantly, this holds for *any* hypothesis h, *any* algorithm Δ, and *any* training set $\mathcal{D}$. Hence, under careful data collection where we know that if the test set is sampled i.i.d. from T, *any hypothesis can be validated based on the observed data only*.

This property, i.e., that we can evaluate the performance of a model without further assumptions on the model itself, is crucial to compare the performance of different methods since different architecture, inference, and hyperparameter choices correspond to different assumptions. Maybe more importantly, this property is also crucial to validate our model assumptions on observed data (given that the sampling assumption holds), since otherwise we could only make statements relative to that our model assumptions hold, which is, of course, much weaker and not informative. Hence, if we need to make model specific assumptions for the validation error to be informative for the generalization error, the train-test paradigm would be relatively meaningless. Of course, the (considerable) challenge is to collect m i.i.d. samples from the *true target* distribution T which can not be guaranteed and is an important assumption on the data collection process.

Table 4: **Estimators and quality measures.**

	Estimators		Risk Measures	
	Monte Carlo	**Horvitz-Thompson**	**Empirical Risk**	**HT Weighted Emp. Risk**
Estimator from $\mathcal{S} \sim \mathsf{S}^m$	$\frac{1}{m}\sum_{x\in\mathcal{S}} x$	$\frac{1}{m}\sum_{x\in\mathcal{S}} \frac{x_i}{p_{\mathsf{T}}(x)}$	$\frac{1}{m}\sum_{x\in\mathcal{S}} \ell(h(x), f(x))$	$\frac{1}{m}\sum_{x\in\mathcal{S}} \frac{\ell(h(x), f(x))}{p_{\mathsf{T}}(x)}$
Estimated expected value	$\mathbb{E}_{X\sim\mathsf{S}}[X]$	$\mathbb{E}_{X\sim\mathsf{T}}[X]$	$\mathbb{E}_{X\sim\mathsf{S}}[\ell(h(X), f(X))]$	$\mathbb{E}_{X\sim\mathsf{T}}[\ell(h(X), f(X))]$



(a) Sampling bias (b) Pareto distribution (c) Degree distribution of MovieLens 100k

Figure 4: **Properties of complex social systems** (a) Graphical model of sampling bias. (b) Illustration of power-law distribution on the example of the Pareto distribution. (c) Degree distribution of MovieLens 100k.

C Complex social systems

In the following, I will discuss how sampling bias and heavy-tailed distributions can occur, and can be connected, in complex social systems. First, *sampling bias* is concerned with how $\mathcal{S}$ is collected. Most standard inference methods assume i.i.d. samples from T, but it is well know that this assumption can be easily violated when sampling in complex systems.

Definition 8 (Sampling bias)**.** Let $\mathcal{S}^t_{ij}$ denote the random variable corresponding to entities $(i, j) \in \mathcal{X}_1 \times \mathcal{X}_2$ j being samples at time t. Samples in complex social systems can then neither be assumed to be independent across time nor independent with regard to the target value f_{ij}, i.e.,

$$P(S^t_{ij}|S^{t-s}_{ij}) \neq P(S^t_{ij}) \quad \text{and} \quad P(S^t_{ij}|f_{ij}) \neq P(S^t_{ij}).$$

Higher arity relations are defined analogously. See also fig. 4a for the assumed sample dependencies.

Sample biases as in definition 8 can be causes by aspects such popularity bias, i.e., if popular items are more likely to be sampled, and quality biases, i.e., if items with higher values for f_{ij} are more likely to be sampled.

A prime example of how sampling bias in the form of popularity bias can lead to power-law distributions, is the influential Barabasi-Albert model [5]. In this model of complex networks, nodes are added to a network one by one and are connected to existing nodes with a probability proportional to their degree, i.e., popularity. Formally, this model is defined as follows:

Definition 9 (Barabasi-Albert model)**.** Let $G = (\mathcal{X}, \mathcal{E})$ be a graph. Furthermore, let $\mathbb{P}(i \sim_t j)$ denote the probability that the edge $i \sim j$ is added at time t to $\mathcal{E}$ and let κ_i denote the degree of node x_i. The Barabasi-Albert model generates then a graph G as follows:

1. Start with a small connected graph G_0 with m nodes.
2. At each time step $t > 0$, add a new node x to G and connect it to m existing nodes in G with a probability proportional to their degree, i.e.,

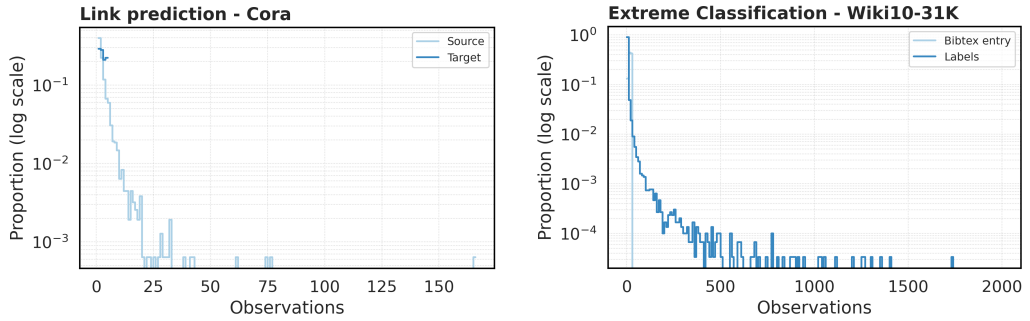
$$\mathbb{P}(i \sim_t v) = \frac{\kappa_i}{\sum_{j \in \mathcal{X}} \kappa_j}$$

It is then well known that definition 9 leads to a power-law degree distribution in G, i.e., a distribution where the probability of a node having k connections is proportional to $k^{-\alpha}$ for some $\alpha > 0$.

D Proof corollary 1 (Necessary condition for test validity)

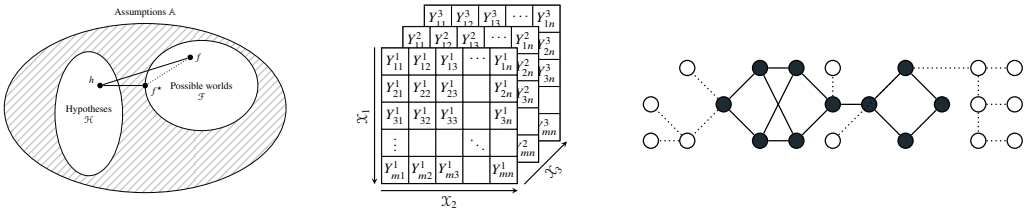
Corollary 1 (Necessary condition for test validity)**.** *Let $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ be an inference setting, let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a positive loss function, and let $\mathcal{H}$ be a hypothesis class. Furthermore, let $\theta \in \mathbb{R}_+$ be any risk estimate for h. Then, if $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ is (ϵ, δ)-test-valid, it must hold that*

$$\exists \mathcal{H}\ \exists h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(L^{\mathsf{T}}_{fh} \leq \epsilon + \theta) \geq 1 - \delta.$$



(a) (b)

Figure 5: **Sample graph degree distributions** for widely used benchmark datasets. (a) Graph learning and link prediction via Cora (b) Extreme classification via Wiki10-31k. As can be seen, all benchmarks follow similar heavy-tailed distributions in their sample graph as the MovieLens dataset in the main text. As such, these benchmarks are subject to the same results and pathologies.



(a) Validity relationships (b) Tensor representation (c) 2-core of a graph

Figure 6: (a) **Relation between assumptions, possible worlds, and hypotheses** Assumptions $\mathbb{A}$ define a set of functions f of which a subset are possible worlds $\mathcal{F}$, i.e., those functions which are also consistent with observations $\mathcal{D}$. While hypotheses $\mathcal{H}$ will often be equivalent to $\mathbb{A}$, they can also be a proper subset of $\mathbb{A}$ and do not need to overlap with $\mathcal{F}$, e.g., due to additional assumptions or computational requirements that constrain $\mathcal{H}$. The functions $f, f^\star$, and h indicate the relevant objects for the necessary conditions in corollary 1 as well as their relationships (solid and dotted lines). (b) **Tensor representation of a function** $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to \mathbb{R}$. (c) **Illustration of the 2-core of a graph**. Nodes within the 2-core are indicated by black. Nodes outside the 2-core are indicated as white, edges that are removed when reducing to the 2-core are indicated as dotted.

Proof. First, note that $|\theta - L_{fh}^\intercal| \leq \epsilon$ is equivalent to $\theta - L_{fh}^\intercal \leq \epsilon \wedge L_{fh}^\intercal - \theta \leq \epsilon$. Furthermore, we have

$$\{f \mid \theta - L_{fh}^\intercal \leq \epsilon \wedge L_{fh}^\intercal - \theta \leq \epsilon\} \subseteq \{f \mid L_{fh}^\intercal - \theta \leq \epsilon\}.$$

It follows then simply from the monotonicity of probability that

$$1 - \delta \leq \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L_{fh}^\intercal| \leq \epsilon) \leq \mathbb{P}_{f \sim \mathsf{F}}(L_{fh}^\intercal - \theta \leq \epsilon) = \mathbb{P}_{f \sim \mathsf{F}}(L_{fh}^\intercal \leq \epsilon + \theta). \qquad \square$$

E Proof lemma 1 (Rank-k underdetermination)

I will first introduce the concept of $\mathcal{S}$-isomerism and connect it to k-connectivity. I will then use these results to proof lemma 1.

First, let $\mathcal{S}_{i,\cdot} = \{j : (i,j) \in \mathcal{S}\}$ denote the set of observed columns for row i and $\mathcal{S}_{\cdot,j} = \{i : (i,j) \in \mathcal{S}\}$ denote the observed rows for column j. Let $\mathcal{S}_{i,\cdot}[\mathbf{F}] \in \mathbb{R}^{m \times |\mathcal{S}_{i,\cdot}|}$ be the sub-matrix of $\mathbf{F} \in \mathbb{R}^{m \times n}$ which is obtained by restricting the columns of $\mathbf{F}$ to the indices in $\mathcal{S}_{i,\cdot}$. Similarly, let $\mathcal{S}_{\cdot,j}[\mathbf{F}] \in \mathbb{R}^{|\mathcal{S}_{\cdot,j}| \times n}$ be the sub-matrix of $\mathbf{F} \in \mathbb{R}^{m \times n}$ which is obtained by restricting the rows of $\mathbf{F}$ to the indices in $\mathcal{S}_{\cdot,j}$. Then, $\mathcal{S}$-isomerism is defined as follows:

Definition 10 ($\mathcal{S}$-Isomeric)**.** Let $\mathbf{F} \in \mathbb{R}^{m \times n}$ and let $\mathcal{S} \subseteq \{1, \ldots, m\} \times \{1, \ldots, n\}$ with $\mathcal{S}_{\cdot,j} \neq \emptyset$. Then, $\mathbf{F}$ is called $\mathcal{S}$-isomeric iff

$$\begin{aligned} \operatorname{rank}\left(\mathcal{S}_{i,\cdot}[\mathbf{F}]\right) &= \operatorname{rank}(\mathbf{F}), \quad \forall i \in 1, \ldots, m \quad \text{and} \\ \operatorname{rank}\left(\mathcal{S}_{\cdot,j}[\mathbf{F}]\right) &= \operatorname{rank}(\mathbf{F}), \quad \forall j \in 1, \ldots, n. \end{aligned}$$

Corollary 3 (Necessary condition for $\mathcal{S}$-isomerism)**.** *Let $\mathcal{S}$ be a sample graph and let rank$(\mathbf{F}) = k$. If $\mathbf{F}$ is $\mathcal{S}$-isomeric, then it must hold that $\mathcal{S}$ is k-connected.*

Proof. First, note that $\text{rank}(\mathbf{X}) \leq \min(m, n)$ for any $\mathbf{X} \in \mathbb{R}^{m \times n}$. Hence, it follows from definition 10 that for a rank-k matrix to be $\mathcal{S}$-isomeric, each column and row needs to have at least k observed entries. Since this is equivalent to k-connectivity, the result follows. □

Lemma 1 (Rank-k underdetermination)**.** *Let $\mathbb{A} = \{f \mid rank(f) \leq k\}$. Then, if $\mathcal{S}$ is not k-connected, the set of possible worlds $\mathcal{F}$ forms a non-empty vector space.*

Proof. Since $\text{rank}(f) = k$ and $\mathcal{S}$ is not k-connected, it follows from corollary 3 that f is not $\mathcal{S}$-isomeric. Hence, it holds via [38, Theorem 3.2] that there exist infinitely many matrices f' that all explain the observed data $\mathcal{S}$ perfectly, i.e.,

$$f' \neq f \quad \wedge \quad \text{rank}(f') \leq \text{rank}(f) \quad \wedge \quad f'_{ij} = f_{ij} \quad \forall (i,j) \in \mathcal{S}.$$

Moreover, it follows from [38, Lemma 5.1] that this set of possible worlds $\mathcal{F}$ forms a non-empty vector space $\mathcal{V}$. □

F Proof lemma 2 (Rank-k test invalidity)

To prove lemma 2, I will first show the following auxiliary proposition:

Proposition 1 (Risk inequality for Bregman projection)**.** *Let $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ be a scalar Bregman divergence and let $f^\star = arg\,inf_f L^\mathsf{T}_{fh}$ be the Bregman projection of h onto a vector space $\mathcal{F}$. Then, it holds that*

$$L^\mathsf{T}_{ff^\star} \leq L^\mathsf{T}_{fh}.$$

Proof. First, note that L^T_{fh} is simply a convex combination of scalar Bregman divergences, i.e.,

$$L^\mathsf{T}_{fh} = \sum_{x \in \mathcal{X}} \ell(f(x), h(x)) p_T(x).$$

Hence, L^T_{fh} itself is a (separable) Bregman divergence. Proposition 1 follows than from the generalized Pythagorean theorem for Bregman divergences [20, Eq. 2.3] since every vector space is a convex set and $f^\star$ is the projection of h onto $\mathcal{F}$, i.e., it holds that

$$L^\mathsf{T}_{fh} \geq L^\mathsf{T}_{ff^\star} + L^\mathsf{T}_{f^\star h} \geq L^\mathsf{T}_{ff^\star}.$$ □

Lemma 2 (Rank-k test-invalidity)**.** *Let $\mathbb{A}$ be identical to lemma 1, let ℓ be a scalar Bregman divergence, let F be the uniform distribution over $\mathcal{F}$, and let T be the uniform distribution over $\mathcal{X}$. Furthermore, let $\theta \in \mathbb{R}_+$ be any risk estimator on a test set. Then, if $\mathcal{S}$ is not k-connected, $(\mathbb{A}, \mathcal{D}, \mathsf{T}, \mathsf{F})$ is test-invalid, i.e., it holds for any $\epsilon > 0$ that*

$$\forall \mathcal{H}\ \forall h \in \mathcal{H} : \mathbb{P}_{f \sim \mathsf{F}}(|\theta - L^\mathsf{T}_{fh}| \leq \epsilon) = 0.$$

Proof. Since the standard uniform distribution is not defined on an entire vector space, I will instead consider the limit of the class of uniform distributions of balls of radius r. Next, since ℓ is Borel measurable, $\mathcal{B}(f, \epsilon) = \{f' \mid L^\mathsf{T}_{ff'} < \epsilon\}$ defines a measurable set around each $f \in \mathcal{F}$. Furthermore, let $\text{Vol}\mathcal{B}(f, \epsilon)$ denote the volume of such an ϵ-"ball". Via the change of variables formula, we know

Table 5: Examples of Scalar Bregman Divergences

Divergence	$\ell(x, y)$	**Divergence**	$\ell(x, y)$
Squared loss	$(x - y)^2$	KL-divergence	$x \log(x/y)$
Log loss	$x \log(x/y) + (1 - x) \log((1 - x)/(1 - y))$	Itakura-Saito	$\frac{x}{y} - \log(x/y) - 1$

then that the volume of $\mathcal{B}(f, r \cdot \epsilon)$, i.e., the volume of the original ball stretched in all directions by $r \geq 1$ is given by

$$\text{Vol } \mathcal{B}(f, r \cdot \epsilon) = \int_{r \cdot \epsilon} dx = \int_{\epsilon} r^{\dim \mathcal{V}} dy = r^{\dim \mathcal{V}} \cdot \text{Vol}\mathcal{B}(f, \epsilon).$$

Next, let U_r denote the uniform distribution over $\mathcal{B}(f, r \cdot \epsilon)$. Then, the probability of sampling a point inside $\mathcal{B}(f, \epsilon)$ when drawing points uniformly from $\mathcal{B}(f, r \cdot \epsilon)$ with $r \geq 1$ is given by

$$\forall f \in \mathcal{F} : \mathbb{P}_{f' \sim \mathsf{U}_r} \left(L^{\mathsf{T}}_{ff'} \leq \epsilon \right) = \frac{\text{Vol } \mathcal{B}(f, \epsilon)}{\text{Vol } \mathcal{B}(f, r \cdot \epsilon)} = \frac{1}{r^{\dim \mathcal{V}}}.$$

Moreover, if $\mathcal{V}$ is non-empty it follows that $\dim \mathcal{V} \geq 1$. Hence, as we increase r to span large parts of $\mathcal{V}$, it holds that

$$\forall \epsilon \forall f \in \mathcal{F} : \lim_{r \to \infty} \mathbb{P}_{f' \sim \mathsf{U}_r} \left(L^{\mathsf{T}}_{ff'} \leq \epsilon \right) = 0. \tag{2}$$

Using eq. (2) I will then show lemma 2 by considering the two cases $h \in \mathcal{F}$ and $h \notin \mathcal{F}$.

If we assume $h \in \mathcal{F}$, lemma 2 follows directly from corollary 1 and $\mathcal{F}$ being a non-empty vector space according to lemma 1 (since $\mathcal{S}$ is not k-connected).

On the other hand, if $h \notin \mathcal{F}$, consider the projection of h onto $\mathcal{F}$ according to ℓ, i.e., $f^\star = \arg\min_f L^{\mathsf{T}}_{fh}$. Since L^{T}_{fh} is a Bregman divergence and $\mathcal{F}$ is a vector space, lemma 2 follows then from corollary 1, proposition 1 and the monotonicity of probability since $L^{\mathsf{T}}_{ff^\star} \leq L^{\mathsf{T}}_{fh}$. □

G Extensions to Higher-arity relations and bounded domains

G.1 Ternary and higher arity relations

First, note that higher arity functions can be represented as tensors of the same order as follows (see also fig. 6b for a visualization):

Definition 11 (Tensor representation)**.** For a function $f : \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_k \to \mathbb{R}$ over *finite* sets of size $|\mathcal{X}_i| = m_i$, we can construct its *tensor representation* $\mathbf{F} \in \mathbb{R}^{m_1 \times m_2 \times \cdots \times m_k}$ via $\mathbf{F}_{ij\ldots,k} = f(x_i, x_j, \ldots, x_k)$ for all $x_i \in \mathcal{X}_1, x_j \in \mathcal{X}_2, \ldots, x_k \in \mathcal{X}_k$.

A trivial extension of theorem 1 and its related results can then be obtained by considering the rank of the projection of the tensor representation of f onto its matrix presentation such that $\mathbf{F}_{ij} = \sum_k f(i, j, k)$ (what equals the assumption that the rank of the predicate mode in f is one). Since predicates in many knowledge graphs are very sparse, the sum over k will often preserve this sparsity and the associated heavy-tailed distributions. In this case, the results of the matrix case extend directly to the tensor case. If this is not the case, it is necessary to extend the matrix analysis of lemma 1 to the tensor case and consider cases where the predicate mode can have rank larger than one. However, this is beyond the scope of this paper and reserved for future work.

G.2 Bounded domains

In many practical applications, the output domain of the function f is bounded, i.e., $\mathcal{Y} \in [0, 1]$. In this case, the main results of this paper can possibly be extended if f can be expressed in terms of a bijective link function g, i.e., $f(x) = g(\phi(x))$ where $\phi : \mathcal{X} \to \mathbb{R}$. Common link functions that are bijective include the logit and probit functions (binary variable) and the log function (Poisson variable). In the following, I provide a brief outline of the argument: Since g is bijective, its inverse exists and we can work on Φ by applying g^{-1} to f. Next, for each column (or row) of Φ, consider the regression problem $\Phi_i = Uv_i$ with the goal of modeling the observed entries of Φ. If the dimensionality (rank) of U and v_i is larger than the number of observed entries in Φ_i, this regression problem becomes an underdetermined system. For areas of the sample graph that are too sparse, we can then find again infinitely many matrices that match all the observed entries of Φ but are different on unobserved entries. The main results of this paper would then extend directly to Φ and via the link function g to the bounded domain. One difference to the results in the unbounded case is that the above argument holds for the degrees in the sample graph, while the unbounded case holds for the stronger k-core condition.

H Inefficiency of scaling and benchmarks

Lemma 1 allows to answer the *scaling* question by asking how many draws from S would be necessary such that all nodes are within the k-core of $\mathbb{S}$ with high probability, i.e., how many samples are needed until arriving at a valid test setting. In the following, I will discuss different ways to approximate this question.

H.1 Scaling and the coupon collector problem

One way to lower bound the number of samples needed to arrive at a valid test setting would be to calculate the number of samples needed to sample each node outside the required k-core at least once. This is an instance of the *coupon collector problem with unequal probabilities*. In particular, let T_k be the number of draws from S until we have collected k distinct nodes from $\mathcal{X}_2$ for an arbitrary node in $\mathcal{X}_1$. Then, it follows from [26, Corollary 4.2] that

$$\mathbb{E}[T_k] = \sum_{q=0}^{k-1} (-1)^{k-1-q} \binom{m-q-1}{m-k} \sum_{|J|=q} \frac{1}{1-P_J} \tag{3}$$

where $P_J = \sum_{j \in J} p_j$ and where $\sum_{|J|=q}$ denotes the sum over *all* subsets J of size q. However, eq. (3) is hard to interpret and for that reason not very useful for our purposes. Moreover, eq. (3) is not even tractable to compute at the scale that we would require for the settings considered in this paper. For instance, assume that we are dealing with a relatively small-scale domain of $|\mathcal{X}| = 10^7$ entities. Since eq. (3) requires to over all subsets of size $k-1$, for a model of rank $k = 10$ this operation alone would require more than

$$\binom{|\mathcal{X}|}{k-1} = \binom{10^7}{9} > 2.75 \cdot 10^{57} \text{ FLOPS}.$$

H.2 Proof for scaling bound in corollary 2

Since the coupon collector problem is not computable, corollary 2 considers an even weaker lower bound and asks how many samples are needed to sample an average node at least once. For a fixed node x_i, this is an instance of *number of trials until first success* and follows a geometric distribution with expected value $T_i = 1/p_i$. Next, for a power-law distribution with $\mathbb{P}(X > x) = u(x)x^{-\alpha}$ it holds that $P(X = x) = u'(x)x^{-(\alpha+1)}$ where u' is also a slowly varying function. Hence, we have

$$T_i = \frac{1}{u'(x_i)x_i^{-(\alpha+1)}} = \frac{x_i^{\alpha+1}}{u'(x_i)}. \tag{4}$$

To illustrate eq. (4), consider the following example using the Pareto distribution to instantiate U'. In this case, we have $T_i = x_i^{\alpha+1}/\alpha x_{\min}^{\alpha}$. For a random node in $\mathcal{X}$ os size n, it holds then that

$$\mathbb{E}_{i\sim\mathsf{U}\{1,n\}}[T_i] = \frac{1}{n}\sum_{i=1}^{n} \frac{x_i^{\alpha+1}}{\alpha x_{\min}^{\alpha}} = \frac{1}{\alpha x_{\min}^{\alpha}} \mathbb{E}_{i\sim\mathsf{U}\{1,n\}}[x_i^{\alpha+1}] \geq \frac{1}{\alpha x_{\min}^{\alpha}} \mathbb{E}_{i\sim\mathsf{U}\{1,n\}}[x_i]^{\alpha+1} = \frac{1}{\alpha x_{\min}^{\alpha}} (n/2)^{\alpha+1}$$

where the inequality follows from Jensen's inequality for $\alpha > 0$. This concludes the proof.

I Experiments

All experiments were computed on a single NVIDIA Volta V100 GPU and implemented using Jax [12], Jaxopt [8], Numpy, and Scipy. All experiments were computed on the MOVIELENS 100k benchmark [27] which is available at https://grouplens.org/datasets/movielens/100k/ and released under a custom license https://files.grouplens.org/datasets/movielens/ml-100k-README.txt.

I.1 Computing possible worlds under rank constraints

To find possible worlds that fit the observed data under a rank-constraint, I will first compute a single subspace for which we can model all observed data with highest accuracy. For this purpose, I am first fitting a matrix $\mathbf{F} = \mathbf{U}\mathbf{V}^\top$ to the observed entries under a rank constraint, i.e., via $\min \|\mathbf{F}_{\mathbb{S}} - \mathbf{Y}_{\mathbb{S}}\|_F^2$ where the constraint $\text{rank}(\mathbf{F} \leq k)$ is enforced simply via $\mathbf{U} \in \mathbb{R}^{m\times k}$, $\mathbf{V} \in \mathbb{R}^{n\times k}$. Next, let $\mathbf{U} = \mathbf{QR}$

be the QR decomposition of $\mathbf{U}$. Then, we know that the set of possible worlds within the subspace spanned by $\mathbf{Q} \in \mathbb{R}^{n \times k}$ must be of the form $\mathcal{P}_{\mathbf{Q}} \cap \mathcal{P}_{\mathcal{S}}$ where

$$\mathcal{P}_{\mathbf{Q}}(\mathbf{M}) = \mathbf{Q}\mathbf{Q}^{\top}\mathbf{M} \quad \text{and} \quad [\mathcal{P}_{\mathcal{S}}(\mathbf{M})]_{ij} = \begin{cases} [\mathbf{M}]_{ij} & \text{if } (i,j) \in \mathcal{S} \\ 0 & \text{otherwise} \end{cases}$$

are the orthogonal projections onto the column space of $\mathbf{Q}$ and the observed entries, respectively. Furthermore, assume that we have already found p matrices that are of rank k and which fit the observed entries $\mathbf{Y}_{\mathcal{S}}$ with high accuracy. We can then find the $p+1$-th matrix by minimizing the following objective:

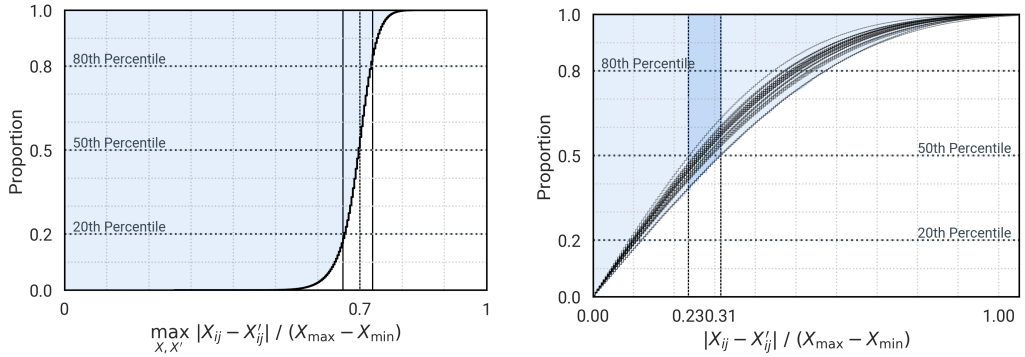
$$\mathbf{X} = \underset{\mathbf{X} \in \mathbb{R}^{m \times n}}{\arg\min} \|\mathcal{P}_{\mathbf{Q}}(\mathbf{X}) - \mathbf{X}\|_F^2 + \|\mathcal{P}_{\mathcal{S}}(\mathbf{X}) - \mathbf{Y}_{\mathcal{S}}\|_F^2 - \textstyle\sum_{i=1}^{p} \|\mathbf{X} - \mathbf{X}_i\|_F^2 \quad \text{s.t.} \quad Y_{\min} \leq X_{ij} \leq Y_{\max}. \tag{5}$$

Importantly, the experimental results in section 3 and fig. 3 hold already for only a *single* subspace $\mathbf{U}$ and considering further subspaces that also explain the observed data can only increase the differences between possible worlds reported in these experimental results.

I.2 Area over the eCDF as expected risk

In the following, I will discuss how the area over the eCDF curves in fig. 3b correspond to the risk $L^{\cup}_{ff'}$ between these pairs of possible worlds. In particular, let E be the random variable corresponding to the normalized absolute error of entries in possible worlds f and f'. Furthermore, let $F_E(x) = \mathbb{P}(E \leq x)$ be the CDF of E. The expected error between both possible worlds (in terms of NAE) is then equivalent to the area *over* the curve of eCDF, i.e.,

$$L^{\cup}_{ff'} = \mathbb{E}_{x \sim \cup}\left[\frac{|f(x) - f'(x)|}{x_{\max} - x_{\min}}\right] = \int_0^1 (1 - F_E(x))dx.$$



(a) eCDF of Maximum NAE

(b) eCDF of Pairwise NAE

Figure 7: **Cora experiments**. Empirical CDF of the maximum NAE (a) and pairwise NAE (b). It can be seen that expected error (area over the curve) behaves similarly to the MovieLens dataset in the main text.

J Related work

In statistics, Meng [43] analyzed a scaling-related question similar to this paper: Given a carefully collected survey with low response rate (small data) or a large, self-reported dataset without data curation (big data), which dataset should you trust more to estimate population averages? For this purpose, Meng introduces an Euler-formula-like identity which connects estimation quality to *data quality*, *data quantity*, and *problem difficulty*. Similar to the results in this paper, Meng shows that data quantity is highly inefficient to overcome issues in data quality, especially sampling related issues. While related in spirit, the results in this paper go beyond the question of surveying and population averages and establish related results in the more general context of inductive inference via formalizing properties of complex social systems and their impact on validity of inferences.

In motivation, this paper is also related to the work of D'Amour et al. [19] who study underspecification of machine learning pipelines as a cause for inference failures. In this context, a machine learning pipeline is "the full procedure followed to train and validate a predictor". A machine learning pipeline is then considered underspecified when it can return many distinct predictors with equivalently strong test performance. This notion of underspecification is closed related to the concepts of possible worlds and validity in this paper.

Srebro et al. [57] studied the problem of matrix completion based on non-uniform observations such as power-laws. However, in contrast to this work, Srebro et al. [57] assume that $\mathsf{S} = \mathsf{T}$. The advantage of this assumption is that it leads to a much simplified learning setting in which valid inferences are indeed possible. However, as I discuss in section 2, I would argue that this is not the problem that many inference settings are concerned with (and that it is questionable in a matrix completion setting as well). Further important results in this context include [49] on low-rank matrix completion from deterministic samples as well as the work of Schnabel et al. [53] and Marlin et al. [40] on learning from biased samples. In contrast to these prior works, I am expanding the setting to the validity of inferences and validation, provide necessary conditions, and situate them explicitly in the context of complex social systems.

K Limitations

As most theoretical work, this work needs to make certain assumptions to make the phenomena of interest amenable to analysis. In this work, the core assumption is that samples in complex social systems follow a heavy-tailed distribution. While this is a very robust finding in social science and widely supported, as discussed in section 2, it limits the results of this paper to this specific setting. For further analysis, this paper further assumes that this heavy-tailed distribution follows a regularly-varying power-law. This is again a supported assumption [64] and allows for a clean theoretical analysis. However, as discussed in section 2, it is still disputed whether samples in complex social systems actually follow this particular form. However, it is undisputed that they follow a heavy-tailed distribution, and as such, while the power-law based results might not apply exactly, their general implications are still supported.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Proofs for impossibility results as well as experimental evidence are provided in section 4.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed throughout the paper and further summarized in supp. K.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: Proof sketches are provided in the main paper, full proofs are included in the supplementary material. All results include a clear statement of assumptions.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The main paper includes an overview of the experimental results. All experimental detail can be found in the supplementary material supp. I. Datasets are publicly available (MovieLens).

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: We intend to release the code for experiments and proofs for the camera ready version. However, at current time, we are not able to provide access to the code. Yet, all experiments should be easily reproducible from the information in the main text and supplementary material.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: See supp. I.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Error bars are not applicable to the experimental analysis of this paper. However, fig. 3b shows the variability over possible worlds.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See supp. I.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research conducted in this paper conforms with the NeurIPS Code of Ethics, i.e., none of the potential harms apply to this work, societal impact of this work is discussed, and impact mitigations are considered as far as they apply.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper is largely theoretical and focused on the validity of the train-test paradigm in complex social systems. As such it has direct connections to *understanding and improving* the social impact of deployed system. While the evaluation of this aspect is beyond the scope of this aspect, I provide discussions of connections to societal aspects such as fairness and participatory data collection. I also discuss fairness aspects in the MovieLens experiments in section 4.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: This paper is purely theoretical and does not release models or data that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The only existing assest used in the paper is the MovieLens 100k dataset for which citations are provided and the license is linked in supp. I.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does currently not release new assets. In the future, code will be released under a CC-BY-NC license.

    Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve study participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.