Is This Correct? Let's Check!

Omri Ben-Eliezer
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

Dan Mikulincer
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

Elchanan Mossel
Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA

Madhu Sudan
School of Engineering and Applied Sciences, Harvard University, Cambridge, MA, USA

Abstract

Societal accumulation of knowledge is a complex process. The correctness of new units of knowledge depends not only on the correctness of new reasoning, but also on the correctness of old units that the new one builds on. The errors in such accumulation processes are often remedied by error correction and detection heuristics. Motivating examples include the scientific process based on scientific publications, and software development based on libraries of code.

Natural processes that aim to keep errors under control, such as peer review in scientific publications, and testing and debugging in software development, would typically check existing pieces of knowledge – both for the reasoning that generated them and the previous facts they rely on. In this work, we present a simple process that models such accumulation of knowledge and study the persistence (or lack thereof) of errors. We consider a simple probabilistic model for the generation of new units of knowledge based on the preferential attachment growth model, which additionally allows for errors. Furthermore, the process includes checks aimed at catching these errors. We investigate when effects of errors persist forever in the system (with positive probability) and when they get rooted out completely by the checking process. The two basic parameters associated with the checking process are the *probability* of conducting a check and the *depth* of the check. We show that errors are rooted out if checks are sufficiently frequent and sufficiently deep. In contrast, shallow or infrequent checks are insufficient to root out errors.

2012 ACM Subject Classification Networks → Error detection and error correction

Keywords and phrases Error Propagation, Preferential Attachment

Digital Object Identifier 10.4230/LIPIcs.ITCS.2023.15

Related Version *Full Version*: https://arxiv.org/abs/2211.12301

Funding *Dan Mikulincer*: Supported in part by a Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826.
Elchanan Mossel: Supported in part by a Simons Investigator Award, Vannevar Bush Faculty Fellowship ONR-N00014-20-1-2826.
Madhu Sudan: Supported in part by a Simons Investigator Award and NSF Award CCF 2152413.

1 Introduction

Understanding the robustness of systems to errors is one of the main goals of theoretical computer science. One set of examples lies within information and coding theory, which study this question for electronic information transmission processes. Another important example is quantum computing; the empirical success of this field crucially relies on the difficult challenge of controlling errors in quantum computers.

In this work we focus on yet another area where errors are prevalent, and error correction has an important everyday role: *societal knowledge accumulation*. Accumulation of knowledge in the modern world is a very rapid yet noisy process, prone to significant errors as new units of knowledge are established [10]. This, in turn, requires proper error mitigation strategies. For instance, the scientific publication process is based upon the assumption that peer reviewing is able to identify errors in submitted papers, ensuring that (for the most part) the scientific literature remains correct and well-founded. This assumption is however very problematic [7, 8]: there are numerous examples of important works with a huge impact on the scientific community, whose findings were later found to be completely incorrect, either because of errors or malicious actions, deeming decades of subsequent research essentially useless. One very recent and prominent example is in the study of Alzheimer's disease [12, 14], where researchers have found evidence that some of the most influential works in the field may be fabricated.

Knowledge accumulation processes such as the scientific process or software development are based on incremental advances that add to the body of knowledge. Each new unit of knowledge may rely on previously discovered ones. Each proclaimed new unit of knowledge may potentially be erroneous on its own or because it relies on an erroneous unit. Without some checks, errors can overwhelm such cumulative processes. As anyone who has ever developed software knows, debugging and testing are crucial for developing reliable code. Similarly, it is unreasonable to trust scientific discoveries in areas where reviews and replication are not taken seriously (see more below).

In these and other areas, natural mechanisms for checking have been introduced. Our work is motivated by the goal of quantifying the success of such procedures: How should such checking mechanisms be measured and evaluated? What are indicators that suggest a proclaimed fact is likely to be true? If one had control of the checking process, what steps would be efficient and effective? Of course, we also do not want to spend too many resources on checking as this slows down the accumulation process, so identifying the sweet spot, where errors are rooted out without spending too many resources on checking, is perhaps the ultimate goal.

In this work, we study these questions under a very simple model. This model which we call the *Cumulative Knowledge Process* (CKP) includes certain ingredients addressing the following fundamental questions (which are essential in any knowledge accumulation process):

1. How is knowledge generated and represented?
2. How do errors arise?
3. When checks occur, what do they check for and what do they do in case an error is found?

We describe the precise model that we work with in Section 1.2. This model involves some choices and here we describe the issues and try to explain our choices. For question (1) above, it is natural to represent the body of knowledge as a *directed acyclic graph* (DAG), where units of knowledge are represented as nodes, and an edge from u to v indicates that v "builds upon" or "inherits from" u. Indeed, since v can only build upon units of knowledge that were created before it, the directed graph describing such relations must be a DAG. Now, when a new node v arrives, we need to pick a subset of "parents" that v shall build upon. The choice of parents should take into account the relevance of the previous node to the new node, which is correlated with latent features such as topics, language, or goals, and also to the importance or impact of the previous node. The former notion (relevance) is somewhat hard to model – multiple models exist and the best choices are still up for debate. The latter is more familiar with models such as preferential attachment forming a good starting point.

In this work, we bypass the challenge of assessing relevance by considering a simpler model where knowledge is represented by a *tree*, i.e., every new node has only one parent. (The challenge of relevance emerges only when we try to model the set of parents, and in particular in determining the correlation between children of two or more parents.) In the tree model, we can bypass this issue and simply consider the setting where a newly generated node picks its (unique) parent based on the preferential attachment model for trees.

Thus in our model, the body of knowledge is a rooted tree, where edges are always directed away from the root and higher degree nodes are more likely to be connected to. We acknowledge that this choice is limiting and more general and realistic DAG processes for cumulative knowledge growth should be studied in future work. (We comment more on this in Section 3.)

We now turn to question (2), i.e, the model of errors. Here we model introduce so-called "primary erroneous facts" in our model by allowing every newly generated node to be an erroneous one with some fixed probability ε. (This is one of three parameters that specifies our process.) Erroneous nodes remain hidden until some checking process reveals the error. Till such stage, erroneous processes continue to produce offspring at the same rate as true nodes. Such "children" nodes and their descendants in the tree are also erroneous and we refer to them as secondary errors.

Finally turning to question (3), i.e., the checking process, we couple the checking of facts with the generation of new children. When a new child is generated, with some probability p a "check" is performed. In this check a path of length k is checked for any primary erroneous node. If any such node is found, all its descendants along the path being checked are "discovered" to be erroneous and no longer participate in the growth.

Our model above captures many natural ingredients of cumulative knowledge (modulo the issue of knowledge being a tree), while still being simple enough to allow for analytic studies. While the error and checking models also involve multiple choices, most are natural (and perhaps not new). The checking process however merits further discussion. First, the model associates checking with the growth of the tree – checks start at new nodes. This, we feel, reflects a natural choice empirically. Units that are not relevant and not used by others are less frequently checked. This is true both in scientific publications and in software development. Indeed, the empirical nature of checking is that facts get checked with probability growing with their impact. In our model, the only impact of a node is in the subtree it generates (and the fraction of the subtree that is not publicly known to be erroneous) and so it makes sense to check only when this impact grows. This leads us to the choice of checking only up to a bounded depth k. Checking all ancestors of a node may make checking too expensive. Our choice ensures that every new node seems to add a "constant" amount of work independent of the size and shape of the tree, making the checking a plausible process.

Some of the basic questions that one might want to study about errors in societal knowledge can be posed in this model. In this paper, we introduce some phenomena one might wish to study, such as: "Does the effect of an error survive for long, or is it fleeting?", and "What would be the characteristics of a CKP that would deem it reliable?", see Definition 1. Our results (see exact formulations in Section 2) can roughly be classified into two types:

- Qualitative results: we identify two contrasting regimes of error propagation. In the first, nodes carrying false information are guaranteed, with probability 1, to have a finite number of descendants. In the second regime, errors may propagate ad infinitum, and there will exist false nodes which serve as roots of trees whose size grows to infinity.
- Quantitative results: within the regimes described above we prove quantitative bounds on the types of error. Thus, even when a false tree grows to infinity we show that, depending on the parameters, its size cannot be too large. Moreover, in the setting where false trees are finite almost surely, we demonstrate a very desirable property: most nodes in the tree carry truthful information.

1.1 Related Work

1.1.1 Noisy Computation and the PMC Model

The question of how to address errors in computation has been extensively studied in many communities. In von Neumann's model of noisy computation [15], the model describes noisy gates and it is shown that with sufficient duplication and provided that the error rate is sufficiently small, errors can be controlled by duplicating gates, see also [4]. The noisy computation model is in some sense stronger than the one considered here as it considers gates with multiple inputs while in our model each unit depends directly only on one unit. However, the noisy computation model has a central planner that can duplicate many copies of the same computation, while in ours such a planner does not exist. Moreover, the noisy computation model has an $\omega(1)$ bigger circuit than the noiseless circuit while the number of operations in our model is only $O(1)$ compared to the model without errors.

Another extensively studied model of error correction that was introduced is called the PMC model [13]. The goal of the PMC model is for functional units to detect the (adversarially) faulty units and this is shown to be achievable under various conditions, see [2] and the references within. Again in the PMC model, it is assumed that the structure of the network can be designed beforehand. Moreover, in our model a node may be incorrect due to errors in previous generations and these could not be checked by immediate neighbors in models of this type.

1.1.2 Local Error Correction and the "Positive Rate" Conjecture

There are related models of "memory" on graphs with noisy gates. The goal of these models is to remember a single bit forever using noisy gates. Again in these models, the errors and the checking are very local (depends just on the immediate neighbors), see e.g., [5, 11].

1.1.3 The Reproducibility and Replication Crisis

There is a large body of work indicating that a substantial fraction of published scientific research is incorrect, as it cannot be replicated or reproduced, see, e.g., [8, 7, 6, 9]. The scientific community is trying to come up with standards and protocols that will reduce that fraction, see e.g., [1]. However, the study of errors in scientific literature mostly ignores the problem of research that is incorrect because it is *based on* incorrect prior research, as these dependencies are hard to understand and control. Our results provide a theoretical framework for addressing this issue.

1.1.4 Knowledge Aggregation vs. Information Spreading

We finally comment on one recent work [3] by a subset of the current authors that considers a similar process where some information spreads noisily through a network with some local checking. In the setting of [3], some node in a given network receives a piece of knowledge and spreads it through the network by local communication. Errors in this setting arise from communication errors when transmitting information, and the checking procedure aims to check the local consistency of knowledge. The paper studied the rate at which potentially erroneous information spreads through the network, as opposed to the spread of the corrected information.

The model in [3] shares similarity with our work in that errors, once generated, may spread and affect other parts of the collective knowledge/belief. However, from this point on, the settings diverge. In particular, a central component of our model is the knowledge graph (the tree) which grows with the process, whereas in [3] the network is extraneous. The checking also models somewhat different settings, and in particular, in our case, when a node is proclaimed to be erroneous, it is actually erroneous (while nodes that proclaim themselves to be true may later end up being found to be false). In contrast, such one-sided guarantees do not hold in the previous work. Finally, the nature of the questions explored in the two models is quite different.

1.2 Models

Here we give the necessary detail and background for our model. The *Cumulative Knowledge Process (CKP)* has states $X_0, X_1, \ldots, X_t, \ldots$ where X_t is given by a finite rooted tree $\mathcal{T}_t$ with each node being given a label from the set $\{\mathsf{PF}, \mathsf{CF}, \mathsf{CT}\}$. We refer to any such labeled tree as a *knowledge state*.

1.2.1 Semantics

CT and CF stand for "Conditionally True" and "Conditionally False" respectively. A node v represents true knowledge if all nodes on the path from v to the root (including v and the root) are CT. We refer to such a node as a True node. All other nodes are False nodes. PF nodes, for "proclaimed false", are those that are publicly false. A priori the CT vs. CF values are not "public" (but can be checked with effort) and so given a node the observable state is either PF or PT (for "proclaimed true") where PT is the observation associated with both labels in $\{\mathsf{CF}, \mathsf{CT}\}$. (Formally the observation is given by the function $O : \{\mathsf{PF}, \mathsf{CT}, \mathsf{CF}\} \to \{\mathsf{PF}, \mathsf{PT}\}$ with $O(\mathsf{PF}) = \mathsf{PF}$ and $O(\mathsf{CT}) = O(\mathsf{CF}) = \mathsf{PT}$).

1.2.2 Parameters

The CKP has three parameters: $\varepsilon \in [0, 1]$ denoting the probability of introducing new errors, $p \in [0, 1]$ denoting the probability of checking, and $k \in \mathbb{N}$ denoting the length of the check. We note that in principle, k may itself be a random variable. For simplicity, we focus on the case where k is constant. However, as will become evident in our results, not much generality is lost by this choice. Given these parameters, we refer to the process X_k as an (ε, p, k)-CKP.

1.2.3 State evolution

Given a state X_t at time t, the state at time $t + 1$, i.e., X_{t+1} is obtained by the following stochastic process:

- **Choosing a parent.** If every node in $\mathcal{T}_t$ is PF then the process "stops", i.e., $X_{t+1} = X_t$. Else first a random PT node u is selected from the tree $\mathcal{T}_t$ with probability proportional to $1 + \deg_{\mathsf{PT}}(u)$, where $\deg_{\mathsf{PT}}(u)$ denotes the number of PT children of u. This is reminiscent of the preferential attachment random tree model. The main difference lies in the fact that once a node has been identified as PF, it may not generate new children. As we shall see, this can drastically change the evolution of the CKP, when compared to preferential attachment trees.
- **Creating a new child.** A new leaf v is attached as a child to u. Set $\mathcal{T}_{t+1}$ to be $\mathcal{T}_t$ with the added leaf v.
- **Error introduction.** v is given the label CT with probability $1 - \varepsilon$ and the label CF with probability ε. Let X_{temp} denote the new knowledge state.
- **Error correction.** With probability p, the path with k edges is "checked" and if an error (either a PF or CF node) is found then all descendants on the path are proclaimed false. Specifically, let $v_0 = v$, $v_1, \ldots, v_k$ denote the path of length k starting at v. (i.e.,

v_i is the parent of v_{i-1}). If all vertices v_i are labelled CT, then $X_{t+1} = X_{\mathsf{temp}}$. Else let j be the smallest index such that v_i is not labeled CT. We modify X_{temp} by relabelling $v_{j'} = \mathsf{PF}$ for every $0 \le j' \le j$. That is, the entire path between v and v_j is labeled PF. X_{t+1} is the resulting knowledge state.

At this point we emphasize the fact that errors can propagate in two ways: either a new CF node is added to the tree in the *Error Introduction* phase, or a CT node is added as the child of a False parent whose observable state is PT.

1.2.4 Initial state

The process we care about starts with X_0 being a single root labelled CT (in other words, the root is always True). This initialization leads to a dichotomy in the tree, False nodes are those nodes which have a CF ancestor, added at some *error introduction* phase, while True nodes are connected to the root by a path of CT nodes. Our aim is to study the difference in behaviors between these two sets.
There is one exception to this initialization rule which we discuss now.

1.2.5 The simple CKP

To build some intuition and to simplify the proofs we consider a simplified version of the (ε, p, k)-CKP process, in which we set $\varepsilon = 0$. We call such a process a (p, k)*-simple CKP*. To avoid trivialities, we always set X_0 to be a single CF root in the simple process. Thus, in this process only CT nodes are added to the tree, all nodes are False, and no new errors are introduced during the process's evolution. This restriction introduces a top-down directionality to the process, since a node may be labeled as PF only after the same happens to its parent.

1.2.6 Phenomena we care about

We now make some definitions to describe the different types of behaviors demonstrated by our results.

▶ **Definition 1** (Properties of CKPs). *For $\varepsilon, p \in [0, 1]$ and $k \in \mathbb{N}$, let $\mathcal{T}_t$ be the knowledge state of an (ε, p, k)-CKP. For a CF node u added to the tree at some time t_u, and for any time $t \ge t_u$, let $\mathcal{T}_t^u$ denote the sub-tree process rooted at u at time t.*

- **Survival of error effects:** *We say that a CKP exhibits* survival of error effects *if the following holds: With positive probability, there exists some CF node u for which the process $\mathcal{T}_t^u$ goes forever. That is, for every t, there is at least one PT node in $\mathcal{T}_t^u$.*
- **Elimination of error effects:** *We say that a CKP exhibits* elimination of error effects *if it does not exhibit survival of error effects. Specifically for every CF node u, there exists a time t after which every node in $\mathcal{T}_t^u$ is PF. In particular, in the simple process, elimination of the error effects means that all nodes have become PF and so the process stops.*
- **φ-reliable process:** *Let $\varphi : \mathbb{N} \to \mathbb{N}$ satisfy $\varphi(t) = o(t)$. We say the CKP process is* φ-reliable *when the following holds. Let M_t stand for the maximal size of a sub-tree in $\mathcal{T}_t$ of False and PT nodes. Then, for any constant $c > 0$,*

$$\mathbb{P}\left(M_t > \varphi(t)\right) \xrightarrow{t \to \infty} 0.$$

Since at time t the tree always has t nodes this condition means that any False *and* PT *node, can only have a negligible number of descendants. We will usually take $\varphi(t) = \Theta(t^a)$, for some $a < 1$.*

- ***δ-highly reliable process:*** *Let $\delta > 0$. We say that the CKP process δ-*highly reliable *if, for any time t, the expected proportion of* False *nodes labeled as* PT*, as opposed to* True *nodes, is at most δ. We say that a process is highly reliable if for every $\delta > 0$ it is δ-highly reliable. (In other words, a process is highly reliable if most of the nodes that declare themselves as* True *are indeed* True*.)*

Observe that all definitions above are only a function of the parameters ε, p, and k.

We note that, other than the fact that survival and elimination of error effects are mutually exclusive, it is not a-priori clear that one definition implies or denies another. Our main results will identify regimes of parameters where the CKP (or its simple variant) satisfies one, or more, of these definitions.

There are additional natural questions one may ask about the model. For example, one could consider highly noisy regimes, when the proportion of False nodes with a PT label is $1 - o(1)$, and so most of the information carried by the process is corrupted. We leave such questions for future investigations.

1.2.7 PT components

We finish this section with the following basic definition, of a proclaimed true (PT) component. This refers to a sub-tree T in any of our processes (CKP or simple CKP), that at some time t consists only of proclaimed true nodes; and furthermore, is maximal with respect to this property.

▶ **Definition 2** (PT Component)**.** *Consider any of the cumulative knowledge processes we define (CKP or simple CKP) at some time $t \geq 0$, and let $\mathcal{T}_t$ denote the (undirected version of the) tree generated by the process. We call the sub-graph of $\mathcal{T}_t$ consisting of all* False *nodes whose observable state is Proclaimed True (*PT*), simply, the* PT *sub-graph. A connected component of the* PT *sub-graph is called a* PT *component. We denote by $\mathcal{C}_{PT}(t)$ the set of all* PT *components in the process at time t.*

If $C \in \mathcal{C}_{PT}(t)$ is a PT *component, we denote by $|C|$ the number of* PT *nodes in C, and if $C' \in \mathcal{C}_{PT}(t+1)$ we use the notation $C' \subset C$ to indicate that C' was created from C in one step. Formally, this means that the root of C' is a descendent of the root of C.*

As we shall see soon, PT components play an essential role in many of our arguments. In particular, we base several potential functions used in our proofs on these components.

1.3 Proof Ideas and Techniques

Our model is based on the preferential attachment model, which has been studied for nearly a century, originating in the work of Yule, [16]. By now, the model is well-understood and the growth and evolution of many quantities of interest have been thoroughly analyzed. Thus, our choice of the model should allow us to tap into many known results, and moreover, the recursive nature of the tree is expected to often lend itself to an exact analysis of the dynamics.

Having mentioned the above, we now remark that the addition of the checking procedure to our model can be thought of as a destructive process, competing with the natural growth process of the preferential attachment tree. While the recursive nature of our process still exists, the dynamics of the preferential attachment tree are often distorted, and some of the underlying symmetry is broken. This becomes particularly evident when one looks at the graph structure of PT nodes, which will now form a forest of fractured components, rather than an ever-expanding tree. Still, we show that the model lends itself to the probabilistic analysis of several relevant functionals, fundamental to our analysis. Most of the functionals are defined as a sum of simpler functionals applied to individual components. We consider functionals that take into account the number of leaves, the degrees of nodes and their depth, the size of components, etc.

The functionals are analyzed by probabilistic techniques in particular using the analysis of (sub/super)-martingales. Roughly speaking, once we show that the processes have a drift in a certain direction, these techniques allow making asymptotic conclusions about the process. Of course, coming up with the right functionals requires creativity and intuition about various aspects of our process.

2 Our Results

We now describe our results. The proofs will appear in the full paper. We begin by addressing the simple CKP model and then proceed by establishing analogs for the general model.

2.1 Results for the Simple Model

Our first two results show that, depending on the parameters p and k, error effects can both survive and be eliminated, in the simple model.

▶ **Theorem 3** (Error effect elimination in the simple model)**.** *For all $p \geq \frac{6}{7}$ and $k \geq 4$, the error effect in the (p, k)-simple CKP is completely eliminated.*

▶ **Theorem 4** (Error effect survival in the simple model)**.** *For all $0 < p \leq \frac{1}{4}$ and $1 < k \leq \infty$, the error effect in the (p, k)-simple CKP survives with positive probability.*

The two theorems above demonstrate contrasting behaviors, which mainly depend on p, the probability of checking for errors. Thus, if one wants to ensure complete elimination of errors, one should be willing to look for errors in a reasonable proportion of knowledge units.

Let us note that the two regimes in Theorems 3 and 4 do not cover the entire parameter space. The behavior of the process for intermediate $p \in (\frac{1}{4}, \frac{6}{7})$ is an interesting question which is left open. In particular, identifying the critical p in which there is a phase transition between survival and elimination of the error effect would be appealing.

To address the role of the parameter k in this model, we focus on the assumption $k \geq 4$ in Theorem 3. Remarkably this is not a technical issue, and the model displays a striking transition when k is small. We show that small changes to the depth of error correction can have dramatic effects.

▶ **Theorem 5** (Error effect survival when $k = 2$)**.** *For any $0 \leq p < 1$ the error effects in the $(p, 2)$-simple CKP survive with positive probability.*

Our next result further elucidates the role of k in the model by examining the reliability of the process. Specifically, we show that even when the error effect in the simple model survives with positive probability, for example, when $p \leq \frac{1}{4}$, the process can still be *reliable*, provided k is large enough.

▶ **Theorem 6.** *Let $p \in (0, 1)$. If $k \in \mathbb{N}$ is such that $\frac{12}{2k-1} \leq p$ then (p, k)-simple CKP is φ-reliable, with $\varphi(t) = \Theta(t^{0.55})$.*

Thus, if k is large, even if a few units of knowledge are periodically checked, it can still be ensured that no false source will corrupt a non-negligible portion of the knowledge base.

At this point, we do not know whether an absolute constant (such as 0.55) is the correct term in the exponent in Theorem 6. In fact, it is reasonable to expect that the actual power will depend on k and perhaps also on p. However, in the full paper we also show that one cannot expect better than polynomial dependency and that with constant probability, there will exist (False) PT components of polynomial size.

2.2 Results for the General Model

In the general model, we first generalize Theorems 4 and 3. Because of the added parameter ε, the overall dependence on the other parameters becomes slightly more complicated. For now, it will suffice to state simplified versions of the results. The reader is referred to full paper for the exact dependencies.

▶ **Theorem 7** (Error effect elimination in the general model)**.** *For every $\varepsilon \in (0,1)$, there exists $p_0 \in (0,1)$ and $k_0 \in \mathbb{N}$, such that for any $p \geq p_0$ and $k \geq k_0$, the error effects in the (ε, p, k)-CKP are completely eliminated.*

▶ **Theorem 8** (Error effect survival in the general model)**.** *For every $\varepsilon \in (0,1)$, there exists $p_0 \in (0,1)$, such that for any $k \in \mathbb{N}$ and $p \leq p_0$, the error effects in the (ε, p, k)-CKP survives.*

Like in the simple model, we see the critical role p plays. At least for low levels of error, one can always invest appropriate effort by performing enough checks and guaranteeing low-term correctness of the information state in the tree. On the other hand, if p is small enough, the process could get overwhelmed by errors.

Using the same ideas used to generalize Theorem 4 into Theorem 8, an analog of Theorem 6 could also be derived for the general model. However, due to the dependencies introduced by new CF nodes, the obtained result is somewhat complicated and not immediately interpretable. Instead, we prove another desirable property that emerges when the error effect is eliminated. Namely, we show that, not only do false sub-trees get eliminated, but that the overall proportion of True nodes in the process remains large. Thus, we show that the process is highly reliable.

▶ **Theorem 9.** *Under the same conditions of Theorem 7, if $p \geq p_0$, and $k \geq k_0$, then the (ε, p, k)-CKP is $O(\varepsilon(1-p))$-highly reliable.*

Note that by a given time t, we should expect to add $\varepsilon(1-p) \cdot t$ CF nodes to the tree; the $(1-p)$ factor comes from the fact that a check is performed a new node is automatically labeled as PF and can thus be disregarded. Thus, $\varepsilon(1-p)$ is the proportion of errors introduced to the tree without accounting for further propagation via attaching new descendants. With this in mind, one way to interpret Theorem 9 is as a statement about types of errors in the process; most errors are expected to come from very shallow sub-trees, and errors, when introduced, tend to be quickly rectified before spreading along the process.

Let us note that there is no hope in improving Theorem 9 by more than a constant. To see this, for a given time $t > 0$, it's enough to consider all the CF nodes added after time $\frac{t}{2}$. It can be shown that with some constant probability, each such node will not spawn a descendent, and hence will survive up to time t. Thus, the expected proportion of CF nodes will be $\Omega(\varepsilon(1-p))$.

3 Discussion and Open Questions

In this paper, we have studied the different behaviors of CKPs when the underlying graph process is a tree. Our results reveal striking differences in the overall shape and persistence of the processes. These differences depend on the interactions between the different parameters.

When k is large, Theorems 3 and 4, for the simple model, along with Theorems 7 and 8, for the general model, identified a phase transition for the error effects, which mainly depends on p. The results establish that the propagation of errors can be completely eliminated, with absolute certainty, as long as we put some constant, which is necessarily not too small, fraction of knowledge units under scrutiny. In contrast, in Theorem 5, we elucidated the dramatic role of the depth,k, and showed that very shallow checks completely nullify the above dependence in p. In particular, when $k = 2$, there is no way to guarantee the elimination of error effects, which should discourage shallow checking procedures.

Other than considering phase transitions, we have also studied the structural properties of the processes, in the different regimes. In Theorem 6 we focused on the case of small p, where the error effects in the simple process can survive. According to the theorem, as long as p is not small, as dictated by k, even when the simple model survives, one may still guarantee that no single error can be connected to most of the entire process. In particular, each surviving component will only have sub-linear size. Finally, Theorem 9, dealt with a regime of the general mode in which error effects are guaranteed to be eliminated. By design, even when error effects are eliminated, the general model continues to evolve, and we show that, by making p still larger we may also guarantee that the proportion of False remains almost minimal.

While we aimed to cover the wide range of possible phenomena exhibited by the CKPs, our work also leaves some open questions. Below we list several such questions and other possible directions for research.

- **Shallowness of checks:** As detailed above, for the simple model, there is a strict phase transition, depending on k. However, our results do not cover the case $k = 3$, and it will be interesting to see whether the error effects can be eliminated in this case.

  ▶ **Question.** *Is there some $p < 1$, such that the error effect in the $(p, 3)$-simple CKP is completely eliminated?*

  It seems that a positive answer to the above question would require a more subtle potential than our exponential potential.
- **Critical parameters:** In a similar vein to the previous question, the, arguably challenging, question of finding the critical p for the transition remains open.

  ▶ **Question.** *What is the value of $p_0 \in (0, 1)$ (which may depend on k), such that for any $p < p_0$ the error effect in the (p, k)-simple CKP survives with positive probability, and for $p > p_0$ the error effect in the (p, k)-simple CKP is completely eliminated?*

  Similar questions are left open with respect to our other definitions. For example, finding the correct polynomial power in Theorem 6 is also of interest.
- **Proportion of false nodes:** Another possible direction would be to to complete the picture presented in Theorem 9, and show a result in the converse direction.

  ▶ **Question.** *Is there a set of non-trivial parameters p, k, ε, such that the expected proportion of False nodes in the (ε, p, k)-CKP is $1 - o(1)$?*

  Note that for True nodes, unlike their False counterparts, there is no a priori probabilistic guarantee on their expected proportion. Thus, it makes sense to study regimes where all nodes, except a negligible proportion, are False. In particular, identifying the possible existence of intermediate regimes where False nodes exist in abundance, yet do not overwhelm the process, is also of interest.

- **More realistic models:** As discussed in the introduction, we considered a simplified model of knowledge accumulation, where each knowledge unit relies on a single existing previous unit. This restrictive assumption naturally leads to the preferential attachment tree we've considered. However, in many cases of interest, trees do not necessarily provide a faithful representation of knowledge accumulation since new knowledge can rely on several different sources, which leads to directed acyclic graphs.

  ▶ **Question.** *Can similar results apply in more general CKPs where the underlying model is a DAG and nodes can have an in-degree larger than* 1*?*

  The crucial point is that any extension of our model to general DAGs must also specify a natural way for a new node to choose a set of "parents". Unlike the tree model, it is not satisfactory to choose a random subset of existing vertices since new units should be more likely to rely on existing units which are similar, in some sense to be defined. So, a general model should define a similarity metric on nodes and choose new parents based on both their degrees and this metric, leading to a more involved analysis.

References

1. David B Allison, Richard M Shiffrin, and Victoria Stodden. Reproducibility of research: Issues and proposed remedies. *Proceedings of the National Academy of Sciences*, 115(11):2561–2562, 2018.
2. Noga Alon, Elchanan Mossel, and Robin Pemantle. Distributed Corruption Detection in Networks. *Theory of Computing*, 16(1):1–23, 2020.
3. Omri Ben-Eliezer, Elchanan Mossel, and Madhu Sudan. Information spread with error correction. *CoRR*, abs/2107.06362, 2021. `arXiv:2107.06362`.
4. William S. Evans and Leonard J. Schulman. Signal propagation and noisy circuits. *IEEE Trans. Inform. Theory*, 45(7):2367–2373, 1999.
5. Lawrence F. Gray. A reader's guide to Gacs's "positive rates" paper. *Journal of Statistical Physics*, 103(1):1–44, 2001.
6. Joseph F. Grcar. Errors and corrections in mathematics literature. *Notices of the AMS*, 60(4):418–425, 2013.
7. John P. A. Ioannidis. Why most published research findings are false. *PLOS Medicine*, 2(8), 2005.
8. John P. A. Ioannidis. Why science is not necessarily self-correcting. *Perspectives on Psychological Science*, 7(6):645–654, 2012.
9. Piers Larcombe and Peter Ridd. The need for a formalised system of quality control for environmental policy-science. *Marine Pollution Bulletin*, 126:449–461, 2018.
10. Helen Longino. The Social Dimensions of Scientific Knowledge. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Summer 2019 edition, 2019.
11. Anuran Makur, Elchanan Mossel, and Yury Polyanskiy. Broadcasting on random directed acyclic graphs. *IEEE Transactions on Information Theory*, 66(2):780–812, 2020.
12. Charles Piller. Blots on a field. *Science*, 377(6604):358–363, 2022.
13. Franco P. Preparata, Gernot Metze, and Robert T. Chien. On the connection assignment problem of diagnosable systems. *IEEE Transactions on Electronic Computers*, 6(EC-16):848–854, 1967.
14. Dennis Selkoe and Jeffrey Cummings. News story miscasts Alzheimer's science. *Science*, 377(6609):934–935, 2022.
15. John von Neumann. Probabilistic logics and the synthesis of reliable organisms from unreliable components. In *Automata studies*, Annals of mathematics studies, no. 34, pages 43–98. Princeton University Press, Princeton, N. J., 1956.
16. George Udny Yule. A mathematical theory of evolution, based on the conclusions of Dr. JC Willis, FR S. *Philosophical transactions of the Royal Society of London. Series B, containing papers of a biological character*, 213(402-410):21–87, 1925.