
Leveraging Benford’s Law for Computational Complexity and Routing Transparency in AI Systems with Cryptographic Timestamps and Event-Triggered Actions

Aaron Goulden

With
Martian & Apart Research

Abstract

This paper proposes an innovative method for embedding computational transparency into AI systems through cryptographic timestamping, Benford’s Law deviation analysis, and event-driven actions. By analyzing the deviation from Benford’s Law in timestamped logs, we derive insights into the computational complexity of query routing and the specific path a query took through a distributed network of trusted, semi-trusted, and untrusted nodes. This method not only enhances the traceability and legitimacy of AI outputs but also introduces the concept of triggering event-based actions, such as compensating analysts or deploying security teams, depending on the sensitivity of the models accessed. We discuss how this system ensures ethical data use, detects potential misuse of AI models, and provides an automated response to high-risk scenarios, offering a transparent and auditable framework for AI operations.

1. Introduction

The integration of AI systems into critical decision-making processes brings with it the challenge of ensuring transparency, accountability, and the ethical use of data. In systems like conversational agents (e.g., ChatGPT), queries can be routed through a network of models with varying trust

levels—trusted, semi-trusted, and untrusted. Each model may have access to different levels of data, and its processing path can affect the quality, cost, and ethical standards of the final output.

This paper proposes a framework that utilizes Benford’s Law deviations in timestamps, combined with cryptographic data, to assess and derive the computational complexity of routing paths. Importantly, we introduce the concept of event-triggered actions: certain model accesses will trigger events that can prompt actions such as compensating analysts, dispatching security teams, or providing alerts when models potentially access sensitive data.

The method provides real-time insights into routing paths, helping ensure ethical AI operation, transparent data use, and efficient allocation of security resources when required.

2. Problem Domain: Trust Levels and Sensitive Data Access in AI Routing

AI systems, particularly those in high-security or ethically-sensitive domains, are often tasked with routing queries through multiple models with varying trust levels. The trust model is typically categorized as:

- **Trusted Nodes:** Models that are fully vetted, secure, and adhere to ethical guidelines. They are given access to sensitive or proprietary data, ensuring privacy and integrity.
- **Semi-Trusted Nodes:** These models are partially vetted and are used for less critical queries or queries that do not involve highly sensitive data.
- **Untrusted Nodes:** Models that are either experimental or have not been fully vetted. These models have restricted access to sensitive data and are generally used for low-risk or exploratory tasks.

In high-security environments, it is crucial to have a mechanism in place that can detect when a query, through its routing, has triggered access to sensitive or unauthorized data. This mechanism should provide immediate action, such as notifying analysts or deploying security teams, depending on the nature of the data and models involved.

3. Benford’s Law for Deviation Analysis

Benford’s Law posits that in many naturally occurring datasets, the first digit of a number follows a logarithmic distribution, where smaller digits (e.g., 1, 2, 3) appear more frequently than larger ones (e.g., 8, 9). This

property can be applied to analyze timestamps in AI logs, providing an effective method for identifying anomalies in the system's behavior.

In this paper, we extend the use of Benford's Law to:

- **Timestamp Analysis:** As queries are processed through AI systems, the generated timestamps are analyzed. These timestamps are expected to follow Benford's Law if the routes taken by the query are typical and involve trusted nodes.
- **Deviation Calculation:** The deviation from the expected distribution of leading digits is calculated for each timestamp. A significant deviation indicates that the query may have traversed an unusual or complex path—likely involving semi-trusted or untrusted models, or models with high computational complexity.
- **Routing Path Insights:** A larger deviation may suggest that the query involved high-risk data access, especially if untrusted models were used. This triggers the need for further scrutiny, potentially leading to the dispatch of security teams or alerts to responsible analysts.

4. Cryptographic Timestamping for Transparency and Event Triggering

The paper introduces a novel mechanism of embedding cryptographic data into timestamps, creating a secure and transparent record for every query processed through the AI system. These cryptographically signed timestamps contain the following information:

1. **Model and Node Identifiers:** Each timestamp encodes the cryptographic signature of the model or node that processed the query. This provides transparency on whether a trusted, semi-trusted, or untrusted model was used.
2. **Data Accessed:** Timestamps also encode a hash of the training data or knowledge base accessed during query processing, ensuring full visibility into what data was used.
3. **Manual Overrides:** If an analyst manually intervenes in the query routing (for ethical, security, or performance reasons), the timestamp includes a cryptographic signature indicating that manual adjustment took place.
4. **Security Events:** The system is designed so that certain models, especially those with potential access to sensitive data, trigger security-related events when queried. For instance, if a model with access to sensitive data (such as PII/PHI or proprietary information) is involved in the query process, the

timestamp will not only reflect the computational path but also trigger an event to:

- a. **Compensate Analysts:** If the model or the query involves an analyst in some capacity (e.g., the analyst manually verifies the query's ethical compliance or its routing), the system can automatically trigger payment to the contractor.
- b. **Dispatch Security Teams:** If the query involves accessing sensitive data (e.g., an untrusted model attempting to process private health information), the system will trigger alerts to deploy security teams to secure the premises or protect the data from unauthorized exposure.

This approach ensures that every action within the system is traceable, and any high-risk action—such as accessing sensitive data or triggering a security event—is flagged and acted upon in real-time.

5. Event-Triggered Compensation and Security Dispatch:

One of the key innovations in this framework is the event-driven response mechanism. By incorporating models that can trigger specific actions, the system not only ensures transparency and ethical data use but also automates response mechanisms for high-risk scenarios.

1. **Analyst Compensation:** When certain models (particularly those with semi-trusted or untrusted status) are queried, or when manual intervention is required for routing decisions, the system can trigger compensation for the analyst. For example, if a model processes a sensitive query involving proprietary or private data, the contractor or analyst responsible for monitoring the integrity of the query can be automatically paid based on predefined contracts.
2. **Security Team Dispatch:** For models with access to sensitive information (such as health data, financial records, or classified datasets), when certain queries trigger access to these data sources, the system can immediately notify the security team. This ensures that there is no data leakage or unauthorized access, and that any potential breach is handled swiftly by deploying physical or digital security teams to safeguard the premises or the system.
3. **Audit and Compliance:** These event-triggered actions are logged in a transparent, immutable ledger, ensuring that all parties involved are aware of the event and can verify the actions taken. The cryptographic timestamps and Benford's Law deviations provide a layer of oversight, making it possible for auditors to confirm that the system responded appropriately to the query's complexity and sensitivity.

6. Implications for Ethical AI Operation and Stakeholder Accountability

The ability to trigger security or compensation events based on model access is critical for maintaining ethical AI operation. This system provides several benefits:

1. **Real-Time Monitoring:** Analysts and security teams can track which models are being accessed, which data is used, and whether there are any anomalies in query routing or timestamp deviations, allowing for immediate corrective actions.
2. **Ethical Oversight:** The system enables continuous monitoring of AI models to ensure they adhere to ethical guidelines. Any access to sensitive data triggers an immediate review or action, ensuring compliance.
3. **Transparency and Accountability:** Cryptographic timestamps provide an immutable audit trail, ensuring that all decisions—whether automated or manually overridden—are transparent and traceable.
4. **Fair Compensation:** The framework ensures that analysts and contractors are compensated for their involvement in high-risk query processing, incentivizing transparency and ethical behavior in the routing and data usage process.

7. Challenges and Future Work

While the proposed system offers a robust framework for AI transparency and ethical compliance, several challenges remain:

1. **Scalability:** Ensuring that timestamping and event-triggered actions can be processed efficiently across millions of queries without significant overhead is a major challenge.
2. **Security:** The cryptographic integrity of the timestamps and event triggers must be ensured, with robust protections against attacks that could tamper with the data.
3. **Model Complexity:** Some complex models may involve dynamic or real-time changes in their behavior. Handling the continuous evaluation of Benford's Law deviations and the complexity of routing in such models is a challenge that needs further research.

Future work will focus on enhancing the scalability of the system, improving the security of timestamp data, and refining the decision-making mechanisms for event-triggered responses.

8. Conclusion

This paper introduces an innovative framework that combines Benford’s Law-based deviation analysis, cryptographic timestamping, and event-driven actions to ensure transparency, ethical compliance, and operational accountability in AI systems. By using timestamp analysis to track query routing and model usage, and triggering security or compensation events when sensitive data is accessed, this framework helps AI systems act responsibly and ethically. It ensures that stakeholders are held accountable for data misuse and provides a mechanism for security teams to respond to potential risks in real-time, marking a significant step toward responsible AI operations.

References

- Benford, F. (1938). The Law of Anomalous Numbers. *Proceedings of the American Philosophical Society*, 78(4), 551–572.
- Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System.
- MacKenzie, D. (2017). *Machine Learning: An Introduction*. Springer.