Research Track

Web mining and content analysis

IPdb: A High-Precision IP Level Industry Categorization of Web Services

Hongxu Chen[1,2], Guanglei Song[†2], Zhiliang Wang[†1,2], Jiahai Yang[1,2], Songyun Wu[1], Jinlei Lin[1], Lin He[1,2], Chenglong Li[1,2]

[1]Tsinghua University, Beijing, China [2]Zhongguancun Laboratory, Beijing, China

清華大學 网络科学与网络空间研究院

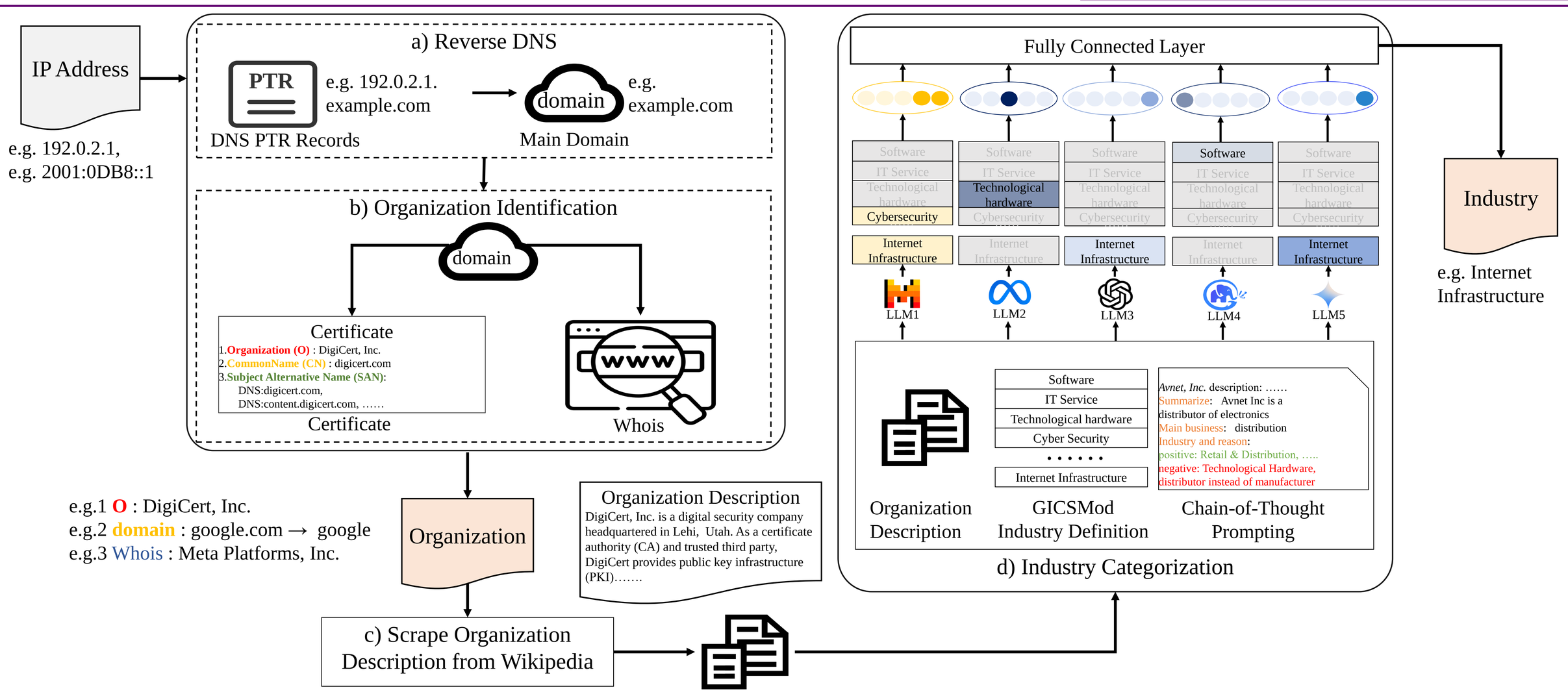
Institute for Network Sciences and Cyberspace, Tsinghua University

Introduction

- **Background:** Web servers linked to critical infrastructure provide critical insights for analysis of **Internet ecosystem** and **cybersecurity**
- **Limitations of prior work:**
  - ✗ Narrow focus on specific organizations limits insights into industry-wide trends or cross-sector security risks
  - ✗ Reliance on AS ownership fails to capture IP addresses leasing
- **Key Contributions**
  - ✓ **High-precision classification**: Evaluation on manually labeled *Gold Standard* dataset: **96% precision, 83% recall**, **94%** fundamentally correct samples (at most one missing label)
  - ✓ **IP-level industry dataset**: 200M+ IPv4/v6 addresses**,** with **200M+ IPv4 addresses** under 12K+ ASes, 90K+ BGP Prefixes, and **746K+ IPv6 addresses** under 1K+ ASes 3K+ BGP Prefixes.
  - ✓ **Critical discoveries★**
- **Impact**
  - Critical infrastructure risk assessment
  - Cross-AS traffic analysis
  - Validation of AS management policies

Methodology of Categorization

- Reverse DNS + Whois/ certificates → identify organizations
- LLM-based classification system (**LLMICS**) using Wikipedia text
  - **Dual confirmation mechanism** and **chain-of-thought technique** to enhance individual LLM's ability
  - **Weighted ensemble framework** to combine multiple LLMs to mitigate LLM hallucination

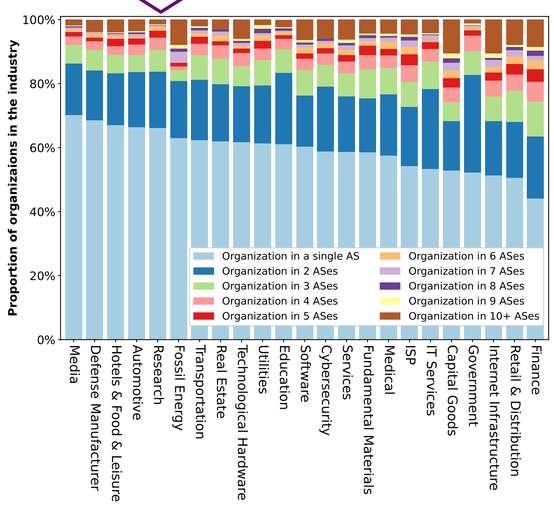


Methodology for Categorizing IP Addresses into Industries

Critical Discoveries★

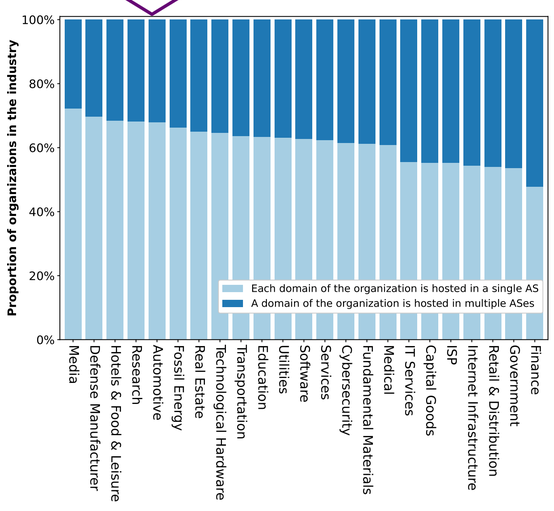
Cross-AS Deployment Prevalence

30–50% of critical infrastructure organizations deploy web services across multiple ASes.



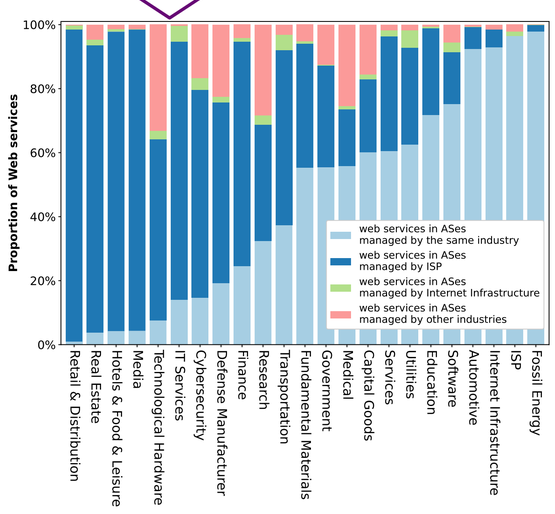
Single-domain dispersion

Cross-AS web service deployment involves both multi-domain distribution and single-domain dispersion (multiple IPs in diverse ASes).



Cross-Industry AS Reliance

When organizations from most industries host their web services in ASes outside their industry, they tend to favor ASes owned by ISPs.



AS-Level Industry Misalignment

Widespread divergence exists between AS ownership and hosted services: misalignment in **87.83% IPv4** and **72.96% IPv6 ASes.**