Is a name enough? A first look into detecting clouds using DNS pointer records

Sousan Tarahomi∗ , Raffaele Sommese∗ , Pieter-Tjerk de Boer∗ , Jeroen Linssen‡ ,
Ralph Holz∗† , Anna Sperotto∗
∗ University of Twente, The Netherlands, † University of Münster, Münster, Germany
‡ Saxion University of Applied Sciences, The Netherlands
{s.tarahomi, r.sommese, p.t.deboer, r.holz, a.sperotto }@utwente.nl, j.m.linssen@saxion.nl

Problem statement

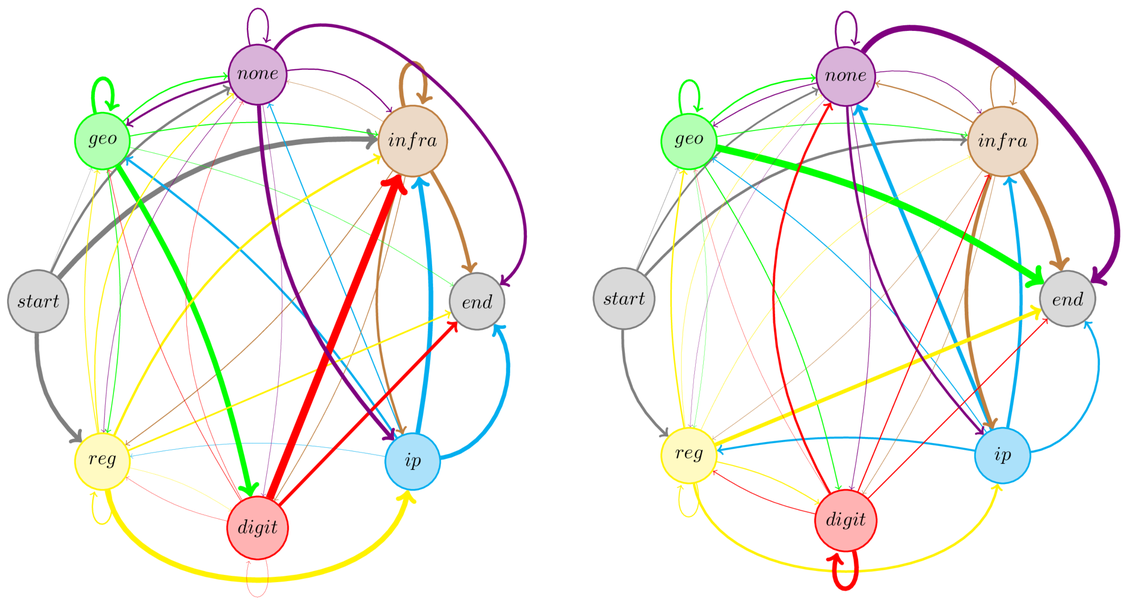
- Understanding **the cloud scope** on the internet is crucial for network management, security, and regulation.
- **Hypergiants cloud providers** often publish the allocation of their network resources, but not the case for smaller providers.
- Despite efforts by **commercial IP intelligence** , there is **a lack of transparency about their identification methods and clarity about its completeness and reliability.**

RQ: How do we distinguish IP cloud from non-cloud ?

Methodology

- We utilized **reverse DNS (Pointer records (PTR)) to** develop a **Markov chain-based classifier** for identifying **patterns and structures** in the reverse DNS naming schemes of cloud providers.

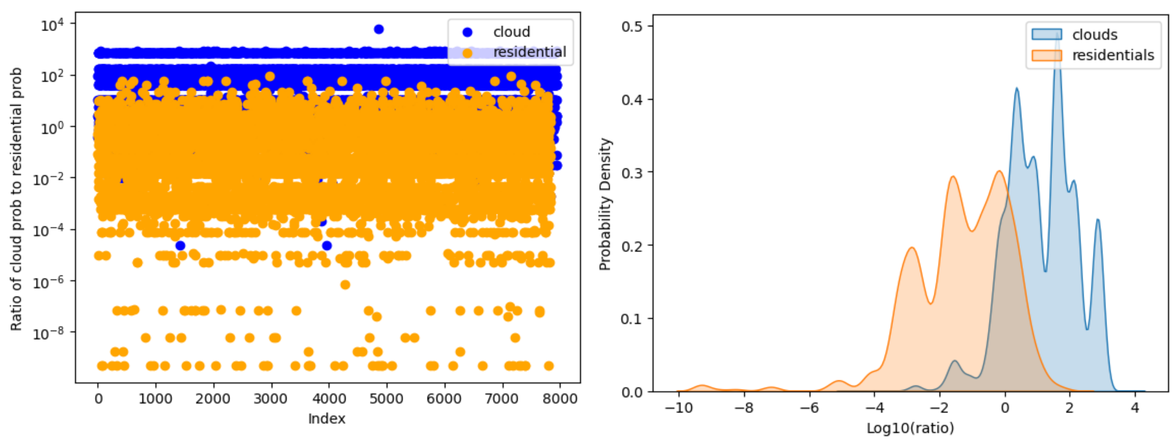
Model



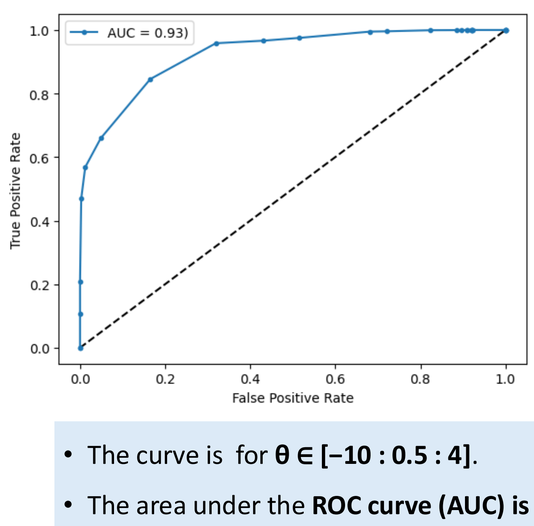
Markov chain model for cloud PTRs **Markov chain model for residential PTRs**

- Each **node i**n these charts represents a **state** in our model and the edges between these nodes show the transition probabilities derived from the training sets.
- The **thickness** of the edges reflects the magnitude of these probabilities.

Results



- These figures show the **higher probability ratio** for cloud samples and a **broader distribution** for residential samples.
- The **overlap** area marks the intrinsic limit of our classification where we can't differentiate between cloud and residential PTR records.



Sequence	Samples in clouds	Samples in residential	Log10(ratio)
(start, reg, end)	197	447	-0.167397
(start, infra, end)	80	110	0.167701
(atart, ip, end)	1	344	-1.554941
(start, none, end)	88	21	-0.137798
(start, geo, end)	9	2	-1.658913

- The curve is for $\theta \in [-10 : 0.5 : 4]$.
- The area under the **ROC curve (AUC) is 0.93 showing high performance for our model.**

- Seven samples of state sequences in the overlap with a **$-3<\log10(ratio)<2$**.
- We are more likely to make a classification error for **short sequences**.

Conclusion

- There are some **common generic sequences**, particularly **single-word PTR names**, shared across datasets, leading to **misclassification**.
- We also identified that **major providers** employ **specific patterns** exclusive to them. These **unique sequences** create **discernible patterns for differentiation**.
- Future research could improve our approach by expanding the **PTR dictionary to include word variants and abbreviations.**

This research was funded by the Centrum voor Veiligheid en Digitalisering (CVD) and by the Netherlands Organization for Scientific Research (NWO) under grant number NWA.1215.18.003 (CATRIN). The research was made possible by OpenINTEL and operational support from the Twente University Centre for Cybersecurity Research (TUCCR).

Find the paper here