SIGCOMM 2008 Review #103A
Updated Friday 29 Feb 2008 8:51:56am PST

Paper #103: Unconstrained Endpoint Profiling (Googling the Internet)

Overall merit: 2. Top 50% but not top 25% of
submitted papers
Reviewer qualification: 4. I know a lot about this area
Novelty: 4. New contribution

===== Paper summary =====

The authors propose to identify what Internet endpoints (hosts) are doing by googling the IP address of each host and evaluating the results. This approach can be used in isolation or to enhance and improved the results from packet and flow level analysis. The overall goal is to categorize the behavior of endpoints according to certain classes such as "Websurfer", "Peer-to-Peer user", or "Gamer".

===== Strengths =====

This paper highlights how much IP address related information is available on the Web and how this information can be used the profile the addresses with an impressive degree of accuracy.

Some of the drawbacks of the approach are discussed in detail.

===== Weaknesses =====

The approach does have potential but the presentation and the comparison with other approaches can be significantly improved.

===== Comments to author =====

First of all I want to congratulate you on a nice idea: to use Google to find out information for endpoint profiling. This is neat and new.

Yet, your approach is based on a lot of assumptions that you somehow still leave implicit:

- First: you seem to find that a single machine is either used for gaming, or for Web browsing, or for ..... Yet, looking at my own Internet use this is certainly not the case. At time x I do a, at time x+t I do b, ...

- Moreover your numbers are certainly not convincing.... Why does your top hit site only have 338 hits in N. America and 395 hits in S. America? This is a very small base. (Table 2). Your overall cache size if very small. Is it statistically significant? Even the results in Table 8 are not convincing, e.g.,

the numbers for gaming are very small (max 261). You have very small hit numbers in Table 2 but some larger numbers in Table 8. How are they related?

- How do you ensure that the data that Google gives you is not out of date? Human behavior changes over time.... Your discussion does not suffice.

- Blinc does not seem to be the best mechanism for classifying endpoints. Did you try any other? An number of unclassified hosts in the order of 70.4% for packet level traces and 97.98% for sampled netflow just does not seem like a good comparison basis.

Overall the approach does have potential but the presentation and the comparison with other approaches can be significantly improved.

Additional comments:

Figure 1 is not readable on some printouts, please make sure it prints on all printers.

Figure 2 needs either a better explanation or another method to visualize the data. I found it hard to correlate the presented data. Perhaps the authors should go for a two-dimensional plot and use different colors for

1.) no Google, no actual endpoint
2.) Google, but no actual endpoint
3.) actual endpoint, but no google
4.) both

and tune the most intense color to the combination you want to stress most.

==========
==
SIGCOMM 2008 Review #103B
Updated Saturday 8 Mar 2008 9:05:35pm PST

Paper #103: Unconstrained Endpoint Profiling (Googling the Internet)

Overall merit: 5. Top 5% of submitted papers!
Reviewer qualification: 4. I know a lot about this area
Novelty: 5. Surprisingly new contribution

===== Paper summary =====

The paper uses information obtained via Google searches as to profile Internet endpoints, enabling application trends and differences between geographic locations to be observed, strengthening traffic classification in packet traces using statistical methods, and more. It is very different than existing methods and works surprisingly well.

===== Strengths =====

A simple idea that works surprisingly well and is new as far as I am aware. It gives some measurement insight into questions such as geographic variation of application usage that were very difficult to tackle previously. This paper should be accepted because it will have a significant impact on Internet measurement.

===== Weaknesses =====

The method could use some more characterization (I would have liked a better characterization of application traffic that the technique does not find well) but this seems minor for a step in a new direction.

===== Comments to author =====

This is a short review, but please do not mistake that for a lack of interest. I thoroughly enjoyed reading your paper for its combination of new idea and execution to demonstrate its value.

The text is somewhat repetitive with the main claims.

In 3.1 near the end, how good is 77%? You might put this in context by saying what the overlap would be if the two were random.

In 3.2, what about the relative ordering of categories? Does your method match the packet trace?

Fig 3, maybe a log scale to show the region of interest more clearly.

In 3.3, can you say more about the 38% of traffic that is missed? This is the main respect in which I hope you will strengthen your paper. Currently, we know that your method is accurate where it finds profile information. I would like to know more about what kinds of application traffic it does not cover well, as we will have to rely on other methods to track that, e.g., Skype?

Similarly, it would be good to quantify the impact of dynamic IP addresses. Do these not affect aggregate studies but mostly preclude host fingerprinting?

==========
==
SIGCOMM 2008 Review #103C
Updated Wednesday 2 Apr 2008 11:54:43am PDT

Paper #103: Unconstrained Endpoint Profiling (Googling the Internet)

Overall merit: 4. Top 10% but not top 5% of submitted

papers
Reviewer qualification: 4. I know a lot about this area
Novelty: 4. New contribution

===== Paper summary =====

Datamining google queries for textual IP addresses in order to profile the characteristics of the ends hosts and those "near" them.

===== Strengths =====

Many people do this on an ad hoc basis, but the idea to doing it comprehensively seems really novel and the results seem fairly encouraging.

===== Comments to author =====

Your model for a profile is binary... you are either X or Y, when in fact many computers are used by multiple people or by people who have different characteristics (at time X they use peer-to-peer at time y they web surf, etc). In fairness, its worth distinguishing between desktop computers and servers... the latter are more likely to have the binary quality you describe, while the former (the majority of machines out there) are likely better described with a mixture model.

In general, dynamic addresses and NAT are going to bite you on the nose. This is especially important since most IP addresses live behind NATs. Can you talk more about this?

In doing a couple experiments I've also found issues with proxies (e.g. like ToR).

Did you pick your categories (e.g. forum user, game abuser) a priori or do they simply reflect the categories you were able to discern from Google. It would be interesting to know if you went in looking for categories that you were not able to identify. I suspect that applications for which there is not a community or a Web-based logging system will not be easy to identify, but I'd like to know if my intuition is correct. Similarly, in 2.2 I'd be nice to know which of these various sources of IP context contribute the most and to what categories.

It would also be nice to know how much work is involved in identifying the keywords used to identify a website class. Does this involve a lot of groveling around at the sites to understand why particular words show up? Is there a reason you can't automate this using a Bayesian network and a bag of words model?

It would be interesting to know how using google compares to a far simpler approach... just taking the IP addresses in the /24, doing reverse DNS and then trying to reach Web servers at those that have names.

blinc is not a great oracle... in fact it can be quite wrong. You should consider doing a comparison with endpoints whose characteristics are known to you personally (i.e. ask a bunch of friends who are gamers to identify their IP addresses to you, etc) and see how you do against a more limited, but more precise, oracle.

The evaluation is not very rigorous. In particular, you don't provide any measurement of statistical significance in your comparisons against blinc.

Is there a relationship between the number of hits in the cache and the quality of your mapping? or are they unrelated and once you have two hits and fetch a URL the quality is as good as its likely to get? To formulate this question as an experiment, what if you only took the URL after the second hit and didn't look at any other urls containing the IP address? Would that matter?

Is there evidence that google's crawling rate is fast enough to deal with use churn in the IP address space?

==========
==
SIGCOMM 2008 Review #103D
Updated Monday 21 Apr 2008 10:43:29am PDT

Paper #103: Unconstrained Endpoint Profiling (Googling the Internet)

Overall merit: 3. Top 25% but not top 10% of submitted papers
Reviewer qualification: 4. I know a lot about this area
Novelty: 5. Surprisingly new contribution

===== Paper summary =====

What cool new information can you find about the online activities of an arbitrary end host in the Internet by simply googling its IP address? This paper argue that it can reveal sufficient information to be able to answer broad questions like "what are the most popular applications being used by clients in different regions of the world?"

===== Strengths =====

The idea of querying Google with IP addresses of random end hosts to infer their potential activities on the Internet is wickedly cool! Easily one of the wildest "out there" proposals I have read in a while.

The authors do an impressive job of mining useful information out of such queries and pointing out potential uses -- e.g., locating active IP address spaces, inferring popular apps etc.,

===== Weaknesses =====

The technique has two fundamental weaknesses:

First, the information that can be gleaned is very (a) high-level, (b) approximate, and (c) limited. The technique gives a lot of information about a lot of things, but it is hard to pin down one thing for which the information is very accurate.

Second, while the results are very intuitive, it is impossible to verify the extent to which inferences are accurate.

Finally, the paper lacks comparison with related work -- for example, how does the information inferred here compare with that gathered and published by web ratings companies like Nielsen.

===== Comments to author =====

I really enjoyed reading the paper. It is well written overall (with the exception of section 2.1.2, which required multiple readings.)

I love the basic idea of googling end hosts, but I have some concerns about the effectiveness of the techniques for profiling end user activities.

It seems to me that your technique while effective at mining lots of information, does not do a specifically good job at inferring specific information.

First, in section 2.2, you mention various likely sources for your information, but what it does not mention is the lack of good sources for a lot of applications, especially P2P apps. For example, I find it jarring that your P2P traffic mix (Table 8) does not contain BitTorrent, by far, the largest P2P app. This is probably because there are no good logs for BitTorrent on the web. You should discuss this limitation in greater detail.

Second, I find your validation (if you could call it that) of your results on application popularity by correlating it with traces (in section 3.2) very hand wavy. All that it seems to say is that the apps that you find from web logs are indeed popular. But, it says nothing about the accuracy of the results. On the contrary, I actually feel that some of your findings might be way off the mark. For example, across the board your results seem to imply that windows OS accounts for significantly less than 50% of all end hosts -- how could this be true? Is this because people using linux tend to use sites or forums that log IP addresses more frequently? If this is the case, how can you be sure of any of your results?

Third, I could not help but wonder how your technique would compare to existing web rating systems that deploy tracers on end hosts to directly measure end host

activities. It would be great to compare the two to evaluate the accuracy of Googling the end point. If I am not mistaken you might be able to get data from neti@home project at Georgia Tech to evaluate how accurate your results are.

Fourth, I find your arguments about how packet-level traces could benefit from this technique rather simplistic (section 3.3). Given that one of the end points of the flow is quite frequently a public server, could not one infer most of this info by simply doing a DNS lookup on the destination IP and port numbers? (As an aside, do you realize that your technique raises severe privacy concerns with sharing network traces with only a few bits of the IP addresses anonymized.)

==========
==
SIGCOMM 2008 Review #103E
Updated Thursday 24 Apr 2008 8:48:21am PDT

Paper #103: Unconstrained Endpoint Profiling (Googling the Internet)

Overall merit: 3. Top 25% but not top 10% of
submitted papers
Reviewer qualification: 2. I have passing familiarity
Novelty: 4. New contribution

===== Paper summary =====

You want to guess what kinds of network activities a given IP address gets up to. Google for it, and automatically classify the hits (gaming, forum, email, &c). Shows some results about how different activities of same IP address tend to correlate, how one can predict what IP address ranges are active, and estimating how local activity is.

===== Strengths =====

An amusing approach. It seems believable that there is lots of information about IP addresses out there.

===== Weaknesses =====

In the end there didn't seem to be much to be learned from this approach. The validation is not convincing. The suggestion that this approach could supplant packet trace analysis seems like a red herring; the two are just different.

===== Comments to author =====

This approach seems like it could be pretty useful, given the right questions to try to answer. But the paper as it stands now seems mostly about the technique, and not very much directed by any compelling questions. Finding a few of the

latter, and re-working the paper to be about investigating them, would make the paper a lot stronger.

The material about traces, throughout the paper, seemed weak. The various attempts at validation in Section 3 were pretty weak; at best, the traces and the UEP results don't completely contradict each other. It might be best to either find a better way to validate, or to present results that don't require validation. Similarly, the claim that UEP can supplant (and is better than) trace analysis aren't convincing. UEP finds a different kind of information than trace analysis; they are not really comparable. In any case, a much better validation would be required. The paper would probably be stronger without the material about UEP supplanting trace analysis.

I wonder if it would be interesting to crawl the web and find out all mentions of all IP addresses. Then you could look at, for example, what fraction of IP addresses are mentioned.

It was very hard for me to guess from the Abstract what the paper was about. A much more concrete treatment would be better. For example, mention some specific questions that UEP can answer, and mention how it works, and mention some specific results. Bringing up traces is confusing; there are too many things you can do with a trace that are not related to this paper (e.g. calculate TCP transfer lengths or window sizes).

2.1.1 should start by saying what the purpose is. Why google for random IP addresses? That doesn't on the face of it seem very useful.

Section 2 should say how the collection of information about IP addresses is driven. Do you crawl the web looking for IP addresses? Exhaustively search Google for every IP address?

Section 3.1 says the goal is "knowing which IPs are active", but in the end you find out something much less specific. It would be good for 3.1 to start by explaining what question you're actually going to answer. It has to do with ranges, not specific IP addresses.

Section 3.2 should show the trace-derived data that it is comparing Table 8 to. As it stands now, 3.2 is a bunch of claims unsubstantiated by data.

It would be great to call out the interesting aspects of Table 8.

4.1.1 should start with a more concrete statement of what you're looking for -- correlation among a given IP addresses activities.