# Respondent-driven Sampling for Characterizing Unstructured Overlays Review

By Yibo Pi and Andrew Quinn

## Citation:

Amir H. Rasti, Mojtaba Torkjazi, Reza Rejaie, Nick Duffield, Walter Willinger and Daniel Stutzbach. Respondent-Driven Sampling for Characterizing Unstructured Overlays. In *Proc. of IEEE Infocom 2009*, pp. 2701-2705, Apr. 2009.

## Summary

This paper presents an approach to characterizing the properties of P2P systems (Gnutella, BitTorrent). They summarize their findings in three ways. First, their new method, Response-Driven Sampling (RDS) outperforms prior approaches in all scenarios. Second, RDS and Metropolized Random Walk (MRW), the previous state of the art, can accurately estimate peer properties. Third, these techniques work in practice.

The authors show that a random walk of the P2P overlay has a fundamental problem in that it will be biased towards nodes which are exceptionally highly connected. RDS is an adaption of random walking which has been proposed in the social sciences and attempts to unbias a random walk. MRW similarly builds on a random walk and attempts to unbias the random walk by modifying the walk to reach each node with equal probability.

Next, the authors do a comprehensive study of the efficacy of these two approaches. They show that RDS outperforms MRW across all static graphs, dynamic graphs and a gathered listing of the graph of the P2P system Gnutella.

# Insights of the paper (“ah-ha” moments)

1. **Sampling instead of full crawling**

A brute force way to characterize an unstructured P2P network is to crawl the full network. However, since P2P networks are large-scale and evolve over time, before a full snapshot of the network can be captured, the network will significantly change. This limits the ability of full crawling approaches to gather accurate network data. Instead of crawling a full network, the author investigates how to sample the network. While the author is not the first one to turn to sampling for this problem, the authors motives for sampling are clearly stated in the paper.

2. **From measuring “hidden population” to unstructured P2P network**

Respondent-driven sampling (RDS) is a variant of snowball sampling, which is proposed to sample hidden populations in social science. “Hidden population” in social science is analogous to P2P networks in the sense that no central directory is available and sampling is spread from existing nodes to others. Snowball sampling achieves unbiased estimation by re-weighting of estimators. RDS follows the similar idea, which re-weights the node properties with the stationary distribution generated by random walk. The results show that RDS is unbiased towards nodes of high degree.

3. **Globally re-weighting to achieve unbiased sampling**

Unlike ordinary or MRW, RDS uses a random walk for sampling and re-weights the estimator using the information of every node traversed by the random walk. Compared to RDS, an ordinary random walk at a node only requires the degree of the node. MRW alters the next-hop selection to achieve uniform distribution, but the next-hop selection relies on the neighbors of the node. RDS re-weights the estimator using all nodes traversed by random walk and thus requires more information than ordinary and metropolized random walk. From the experimental results, we can see that the use of global information helps RDS outperform MRW in hierarchical scale-free graphs. Since RDS does not rely on the degree of nodes, it can easily walk out clusters with well interconnected low degree nodes, while MRW walker has low probability of traversing from a low degree node to a high degree one.

4. **Parallel Sampling to reduce required walk length**

If only one walker is used at a time, hundreds of steps are required to achieve a high accuracy. Walking hundreds of steps is time-consuming and is not effective in dynamic graphs. The author suggests using parallel walkers and discusses their tradeoffs. In parallel sampling, walkers all start from the same node, which increases redundant sampling of nodes around the starting node and reduces the accuracy. The experiments show that in dynamic graphs, the performance of both RDS and MRW degrades after a certain walk length. Meanwhile, RDS and MRW perform better when the session length is longer, because a shorter session length implies a more dynamic network.

5. **Accuracy of RDS and walk length follow power law**

It is interesting to note that in the experimental results, if we consider the accuracy of RDS as probability, the distribution of walk lengths follows power law. In the experiment settings, the node degree distribution is set to be power-law. It is possible that the power-law distribution of node degrees is the reason why the distribution of walk lengths is also power-law.

## Limitations of the paper

1. **Parallel sampling starting from a single point**

In parallel sampling, the accuracy of RDS first increases with walk length, but the accuracy begins to decrease when the walk length is larger than 100. The degradation of accuracy is because all walkers start from the same node, resulting in redundant sampling of nodes around the starting point. Since a larger walk length is not always desirable, it makes parameter tuning difficult. A possible improvement for RDS is to start walkers at different nodes.

**2. Unknown coverage of sampling**

We do not have clear understanding about the coverage of random walk (i.e., the percentage of nodes and edges that are sampled) in order to achieve a certain accuracy. Even though the author showed the performance of RDS with respect to walk length, walk length does not tell us the coverage of random walk. It is possible that the walk length is large, but the coverage is small.

**3. Re-weighting for parallel sampling is unclear**

The re-weighting procedure for a single walker is clear, but how to re-weight the random walks over parallel walkers is not discussed in detail.

**4. Latencies between peers are randomly selected**

In the experiments, latencies between peers are randomly selected from the King data set, which means the latencies from a node to its peers are random. This may not be the case for some nodes. If this is not the case, will the performance of RDS on latency distribution be affected?

**5. Why can cruiser collect complete snapshots?**

In the last experiment on evaluating over Gnutella, a method, cruiser, is used to collect complete snapshots. However, the author did not mention the disadvantages of cruiser and why cruiser works for a full crawl. If cruiser is a good method for doing a full crawl, why do we even need sampling?

## Avenues for future work

One major issue in these papers is figuring out three properties about the techniques and measurements from the paper:

1. Can these techniques presented by the paper be improved?
2. Are these techniques likely to be useful for the envisioned domain?
3. Can these techniques be applied to other domains?

First, it is likely that the techniques presented by this paper be improved. For example, the authors allude to the idea that the techniques in RDS and MRW may be complimentary; but do not investigate this claim at all. Additionally, the parallel sampling can be improved by starting walkers from different nodes. However, the results from the paper are ultimately quite strong; it isn't clear how much there is to gain from improving the techniques in the paper.

Second, are the tools created by the paper useful for the envisioned domain. I.e. Are we likely to see RDS, or MRW for that matter, used in order to understand the nature of P2P systems any time soon. Our position is decidedly no. It is a little bit challenging to get reliable scientific measurements on the distribution of Internet use, but we do our best nonetheless. Variety(http://variety.com/2015/digital/news/netflix-bandwidth-usage-internet-traffic-1201507187/ ) uses data from Sandvine, a broadband company, which reports that BitTorrent accounted for only 4.3% of aggregate internet traffic in 2015. One of the major selling points of this work was how important these P2P systems are, not just because of their use but because they account for enough internet traffic that we need to understand them to understand the internet. I'm not sure that this argument holds water when BitTorrent accounts for only 4.3% of traffic. P2P systems may benefit from the performance that these approaches provide, but it is unlikely to matter much in the broader vision of improving visibility into how the Internet works.

Finally, can these techniques be useful in other domains? These techniques can be useful anytime two conditions hold. First, there is an overlay network. Second, nodes in the network are not coordinated by a central authority (i.e. no one entity knows the state of all nodes). This can happen either because they are owned by different entities or because they are transient in some way (the devices enter and leave the network freely). Above, we argue that traditional P2P systems are dying; the world of overlay networks as described by this paper is likely dwindling. However, sensor networks clearly have uncoordinated and transient devices, where individual sensors may disappear and reappear in the network depending on a variety of environmental factors. In sensor networks understanding the connections within the network at any one time is likely to be extremely useful. Another promising area is that of pervasive computing, where computers are becoming more used in our daily lives.  While pervasive computers of today (smartphones, smart watches, etc.) do not often communicate with devices owned by other people, there are many potential areas of pervasive computing which are likely to involve cross owner communication in the future, such as autonomous driving. The work on understanding overlay networks in this work can likely be applied to these areas, as pervasive devices are not coordinated by a central authority.