WalkAbout, a Random Walk Based Framework to Characterize Online Social Networks

Reza Motamedi
Computer and Information science Department
University of Oregon
motamedi@cs.uoregon.edu
DRP Report

December 6, 2012

Abstract

The structure of networked systems such as online social networks (OSN) or power grids is often represented with a graph. Characterizing the connectivity features of such a graph reveals the structural properties of the corresponding system. In particular, identifying tightly connected regions (i.e. clusters) in a graph provides a valuable insight about its connectivity structure. However, existing cluster detection techniques often require the entire graph, only detect clusters with a relatively small diameter and do not gracefully scale to large graphs.

This project presents a new technique, called WalkAbout, for identifying tightly connected regions in large graphs using short random walks. The key idea is that the ratio of visit (by short random walks) to degree for individual nodes can be used to distinguish nodes in different regions of a graph. We describe the technique, its key features and limitations and perform several validations. Using WalkAbout, we characterize regional connectivity of four large OSNs and demonstrate that the structure of these systems composes of a couple of pronounced regions. Finally, we explore the underlying causes for the formation of these regions in our target OSNs.

1 Introduction

Networked systems range from the world wide web, the Internet router topology and the power grid through online social networks. These systems have structures that can be represented with graphs. Characteristics of these systems can be studied with the analysis of large graphs representing these systems. Thus, graph analysis metrics are extensively used to understand properties of the underlying system that is represented by this large graph.

Social ties in social networks are often represented in the form of a large graph. Data collection was always a major hurdle in conducting analysis on large-scale social systems. Traditional social systems were captured by questionnaires, which proved to be a slow process. In the past few years, the booming growth of online social networks provided scientists with a new source of information at their fingertips. Data collection from remote online websites with limitations enforced by their administrators on the number of profiles that can be viewed per time unit is still a major challenge in study of online social networks. Yet it is still more scalable than traditional methods of capturing social relationships. Thus, large graphs of online social networks have extensively been analyzed and studied in the past few years to reveal properties of these social networks and to show how social behaviors and interactions lead to the formation of social structures that these graphs present.

Prior studies have shown that graphs of online social networks, regardless of their era and area exhibit similar properties. For instance, they all have a small diameter, high level of clustering and their degree distribution follows power low [1]. The study of social graphs yields significant information about social behaviors of systems that the graph represents. For instance, high clustering coefficient of social networks uncovers how links which represent social ties are added to network between users with common friends. Studies of large online social graphs have yet another important contribution since they provide valuable information on how these online networked systems behave and evolve. This information can then be used to ease maintenance of the system and to augment the system with new features that can attract more users.

Clusters of tightly connected nodes in a graph provides a valuable insight about graph's connectivity structure. Clusters can be used to produce a low resolution view of the graph. In this view a graph is summarized by characterizing its clusters and interconnection of its clusters. Social graphs are known to have clusters of tightly connected nodes, which is a result of their growth by evolution along social and foci[1] ties [1]. These tightly connected clusters are inter-connected by sparse bottlenecks. In this study, we aim to produce this low resolution view of the graph and focus on the connectivity properties of large graphs via characterizing connectivity features of tight connected regions in these graphs.

We propose a technique, called WalkAbout, to capture regional connectivity features of a graph using random walks. To this end, we use local properties captured by many short random walks to extract tightly connected regions in the graph. Random walks are widely known to stay in a tightly connected cluster [2] because the likelihood of random walk escaping from a set of connected nodes is proportional to the number of edges internal to the set compared to the number of edges connecting the set to rest of the graph. We argue that the collection of random walks that are stuck in a region can be used to extract properties of that region. Thus, regions are extracted by clustering nodes based on their observed properties. WalkAbout then measures regional connectivity using a random walk in steady state to give a full coarse view of the graph.

WalkAbout is scalable and solely relies on applicability of random walks on the graph. As long as random walks can be performed, WalkAbout can be used to extract regional connectivity features of the graph. In addition WalkAbout does not rely on availability of the entire graph's snapshot. Random walks can be performed in an online manner, which eliminates the need for availability of the whole graph.

In short, WalkAbout

- Examines transient phase of short random walks looking for clues of existing regions. This examination reveals if graphs regions are detectable for certain length of short random walks.
- Extracts regions by clustering nodes based on different properties exhibited by nodes of each region. WalkAbout specifically uses visit-degree ratio as a signal to detect regions.
- Having detected the regions, it evaluates regional connectivity by capturing random walk transitions between regions.
- Produces a regional summary of graph, using its regions and their properties.

This paper is organized as follows. Section 2 portrays the state of the art studies on social networks and graph clustering algorithms. Section 3 and 4 presents the rationale behind our proposed technique and provides evidences on its correctness. Section 5 briefly explains WalkAbout methodology and its parameters. In Section 6, Section 7 and Section 8 we introduce our target OSNs, apply the WalkAbout methodology and finally represent the regional view which is the out come of the WalkAbout. Section 9 thoroughly investigates link level connectivity of regions and shows qualities of detected regions by measuring the number of links that connect nodes of a region to other nodes of the same region and nodes of other regions. In Section 10 we investigate root causes that lead to formation of regions using group membership information that discloses common interest among users and we finally conclude this paper in Section 11.

[1]Social and foci ties are connections linking users to their friends and users to their interest respectively.

2 Related Work

Random walk based analysis of graphs were among the earliest techniques to study large graphs. Random walks are extensively used in research to extract clusters in graphs. Techniques to detect a local tightly connected cluster surrounding a node mainly use random walks. Different Sybil defense mechanisms like SybilGuard [3], SybilLimit [4] and Sybilinfer [5] in online social networks use random walks to detect clusters of tight connected nodes in social graphs which might have fake identities. One shortcoming of these methods is their local nature. While these methods are capable of finding clusters of tightly connected node surrounding a seed node, they cannot be used to globally cluster the graph.

Random walk based techniques were also proposed for global cluster detection. It was suggested by [6] that random walk traces on complex networks reveal community structures. They have introduced an information theoretic approach that reveals community structures in weighted and directed networks, to comprehend the multipartite organization of large-scale biological and social systems. Harel et al. in [2] propose an approach to detect clusters of a graph based on deterministic analysis of random walks on the weighted graph. They deterministically measure the probability of traversing links by random walks. They use the probability to locate bottlenecks in the graph. One of most novel work in study of graph was done by Maggioni et. all [7]. In his work he uses the random walk distance of nodes to detect clusters of tightly connected nodes by applying the diffusion wavelet transform on these distances. He then extends his work by providing multi-resolutions view of the graph. The major shortcoming of random walk based techniques for global clustering is scalability.

Spectral clustering techniques and their approximations have also gained great amount of attention. Using algebraic representations of graph, these methods evaluate tightly connected clusters by computing eigen vectors of graph's the Laplacian matrix [8] [9] [10]. In these methods graphs are hierarchically divided into two or more sparsely connected clusters, until a further split does not result in well-separated clusters. Scalability and feasibility are two major concerns since evaluation of eigen values and eigen vectors are computationally expensive.

WalkAbout also relies on random walks and the fact that short random walks remain in the same neighborhood as their starting point. It uses visit-degree proportionality as a signal to detect all regions in the graph. A collection of random walks in a region have visit-degree proportional to the region's density. Thus it can detects all tightly connected regions that have different densities. As a result WalkAbout is largely scalable and easy to deploy on large scale graphs.

3 WalkAbout - Rationale

In this section, we establish the background and rationale behind the intuitions used to devise WalkAbout. Section 3.1 contains the background on the theory of random walks as well as our assumptions and evidences on correctness of them. Section 3.2 introduces the *visit-degree phenomenon*. This phenomenon is the foundation of our methodology to characterize graph connectivity. Section 3.3 presents an intuitive model to explain the observed phenomenon in Section 3.2 and its contradiction with intuitions derived from theoretic background of Section 3.1.

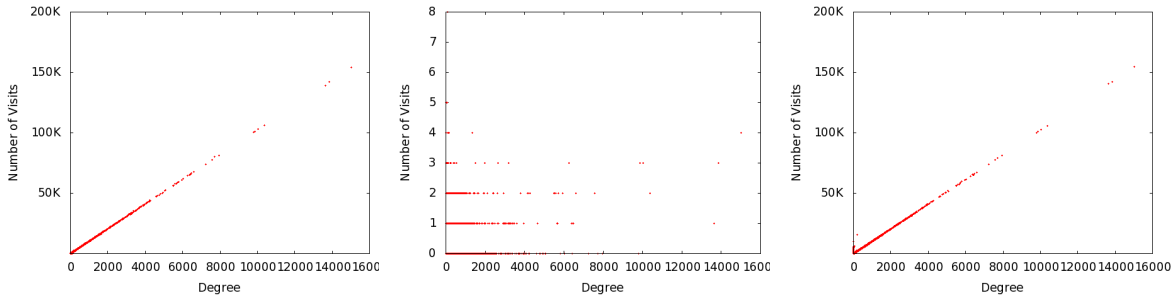
3.1 Background

A *random walk* is a stochastic process on a graph. Given a graph and a starting node, a random walk selects a neighbor of the node it is currently visiting at random at each step. Given a graph $G = [V, E]$ (V and E denote set of graph's vertexes and edges respectively), the theory of random walks [11] suggests that having a random walk in its steady state, there exists a proportionality between the number of times node x is visited by the random walk and the degree of the node. In other words

$$\forall n \in V : \frac{visits(x)}{degree(x)} \sim \frac{walk\ length}{2|E|} \tag{1}$$

Figure 1(a) shows a plot of visits vs. degree for a random walk in its steady state on the graph of LiveJournal [2]. This figure depicts the number of visits vs. degree of nodes visited by random walks. The length of the random walk in this figure is 10^9 steps. The figure evidently shows a constant visit-degree ratio for this walk length. Decreasing the length of the random walk compromises the visit-degree proportionality. Figure 1(b) shows plots of visits vs. degree for walk length of 10^4 steps. As the figure shows, the number of visits and degree are not proportional for this walk length.

Figure 1(c) shows the plot of the aggregate number of visits vs. degree, collected from 10^5 random walks, each with length 10^4. Interestingly, this experiment shows that although visit-degree proportionality does not hold for each random walk as in Figure 1(b), the aggregate visit count and degree are proportional. Thus Figure 1 shows that a random walk on LiveJournal in its steady state can be replaced by multiple random walks of length 10^4 steps.



(a) Steady state (b) Single long walk (c) Multiple long walks

Figure 1: Comparison of visit count vs. degree plots. 1(a): Single random walk with length 1,000,000,000 steps on LiveJournal OSN, 1(b): Single random walk with length 10K on the same graph, 1(c): Same budget as 1(a) spent on 100,000 random walks each with the length of 10,000 steps

Intuitively the number of visits by the random walk depends on both the walk's starting point and the degree of the node. Among two nodes with the same degree, a random walk has the higher chance of visiting the one residing in the neighborhood of its starting node. Thus, using the short walk length, plot of visits vs. degree in Figure 1(b) does not show proportionality of visit and degree either because *i*) The short length of walker makes the visits degree ratio noisy for this walk length, or *ii*) The effect of starting node prevents the walker from showing the similar visit-degree ratio.

We believe that the short length of each random walk prevents it from exhibiting the visit-degree proportionality only because of its small number of credits it can distribute as visit counts. Since the aggregate visits of nodes matches the visits count of the walk in steady state, we believe that for each random walk number of visits does not depend on its starting point.

3.2 Visits vs. Degree Phenomenon

Random walks have the tendency to stay in a tightly connected region. Shorter random walks have the higher chance of staying in the neighborhood of their starting point. For a short random walk, the effect of the

[2]LiveJournal snapshot is used in Sections 3 and 4 to present some preliminary evidence of the correctness of our intuition. This graph will be completely introduced in Section 5.

starting node is an important factor on the number of visits to a node. In this section we decrease random walk length and repeat the previous experiment. Figure 2 shows plot of visit vs. degree for random walks with the length of 100 steps on LiveJournal. As figure shows, visits-degree proportionality we observed in Figure 1 does not exist when we use random walks of length 100 steps. What this figure simply presents is similar degree nodes with different number of visits. There are two interesting details in this plots that is worth mentioning:

1) $10,000,000$ walkers each with length 100 do not show similar visit-degree ratio to the random walk in steady state, but $10,000$ walkers each with length $10,000$ do.

2) Although all nodes do not have similar visit-degree ratio, nodes can be divided into subsets for which there exists a proportionality of visit-degree. In other words, instead of lining up along one angle, nodes are lined up along multiple angles in the visit vs. degree plot.

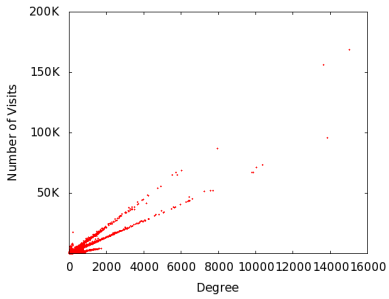


Figure 2: Lack of visits count and degree proportionality for 10M walkers each with length 100 on LiveJournal

To explain multiple dominant visit-degree ratios we use the notion of random walks that are stuck in a region. Random walks have the tendency to stay in the same neighborhood as their starting node. Having a region of tightly connected nodes, the collection of random walks that stay in this region can replace single random walk that only visits nodes of this region. Thus this collection shows a visit-degree ratio for that specific region. Since graphs of social networks contain multiple loosely connected regions, each of these regions represents itself as one the angles to which nodes are aligned in scatter plots. Increase in length of random walk, reduces number walkers that stay in a single region, hence visit-degree ratio exhibited by all walkers reveals connectivity features of the whole graph instead of graphs' regions.

3.3 Intuitive Model

The goal of this section to formulate our intuitive explanation of Section 3.2. To this end, we propose a model to explain why visit-degree ratio is not equal for all nodes.

Suppose $G_1 = [V_1, E_1]$ and $G_2 = [V_2, E_2]$ are two disjoint graphs. We can consider these as a single graph with two disconnected components. If we run n long random walks with the length of t, visit-degree proportionality of long random walks suggests that each of these connected components have a distinct proportionality, i.e. the number of visits of node x with degree $degree(x)$ in connected component c is proportional to $\frac{degree(x)}{2\times E_c}$. This results in same degree nodes in different components to have different visit count since properties of their host components are different. In other words, total number of visits depends on the number of random walks that start from node n's component and proportional visit count it receives in that component. Equation 3 estimates number of visits of node n.

$$visits_n = (n \times \frac{|V|_c}{|V|}) \times (t \times \frac{degree(x)}{(2 \times |E|_c)}) \tag{2}$$

$$= n \times t \times degree(x) \times \frac{|V|_c}{(2 \times |E|_c \times |V|)} \tag{3}$$

In the above formula, $n \times |V|_c/|V|$ estimates the number of random walks which start from the node's component and $t \times degree(x)/(2 \times |E|_c)$ estimates the number of visits to that nodes by a random walk with t steps in the component c. Note that $\frac{|E|_c}{|V|_c}$ is the average node degree of component c. We can see that the number of visits in this example is proportional component's density or its average degree. This dis-proportionality will exhibit itself as two separate lines along which nodes are lined up in visits vs. degree scatter plot.

If we add a few edges to connect two components and make a single connected component, we still have implicit regions created by tightly connected groups of nodes. Based on the hypothesis of random walks getting stuck in more clustered areas, intuitively we can still see different visit-degree proportionalities for region. Of course, adding the bottleneck to the graph gives the chance to some walkers to reach both components. But sparseness of bottleneck limits the number of walks that may cross the bottleneck.

It worth mentioning, if average degree of two components were similar, visit-degree ratios of nodes in those components would be similar values. Hence this signal can be used only for graphs that are comprised of regions with different average degrees.

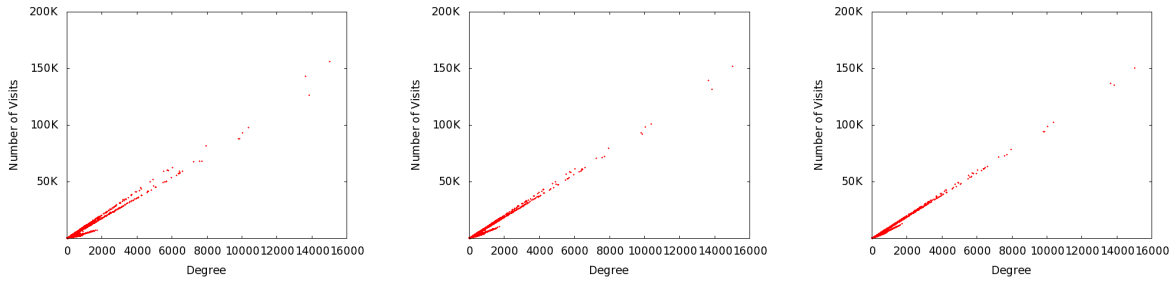
4 Preliminary Model Validation

Our technique in capturing the connectivity features of large graphs from visit-degree ratio is built upon this hypothesis that short random walks remain in tightly connected areas of the graph around their starting point. Thus each short random walk contributes to visits of nodes in the same a region of the graph as the region of its starting node. A collection of random walks that start from a region perceive a that region's properties instead of properties of the whole graph. Thus observing multiple local proportionalities in visit-degree ratio instead of a single one is resulted from different properties of different regions of the graph. In this section we investigate the correctness of our hypothesis and hence our intuitive model.

4.1 Validation with Randomizing The Graph

In our first attempt in unveiling the connection between different visit-degree ratio and connectivity features, we increase width of the bottlenecks connecting different areas of the graph. To this end, we apply a degree preserving randomization method on the graph of LiveJournal. Randomization algorithm simply swaps one end of two randomly selected links from the graph. We can tune this method by limiting percentage of links in the graphs that are altered as the result of its application. Randomizing a graph simply tears down its local connectivity structures by replacing local links with random ones. As a result, internal connectivity of region decreases and the width of bottlenecks increases.

Figure 3 shows visit vs. degree plots for different randomized versions of the LiveJournal each with a different level of randomization. Figures show that as randomization increases the width of bottleneck connecting graph's regions, alignment angles of nodes become closer in scatter plots. Wider bottlenecks give random walks a higher chance of escape from the starting region. Thus, increase in the number of walkers crossing the bottleneck, reduces the distinction of regions due to the gap between their alignment angels.



(a) Randomizing 10% of edges (b) Randomizing 25% of edges (c) Randomizing 50% of edges

Figure 3: Comparison of visits per degree phenomenon after randomizing different percentage of edges

4.2 Validation with Synthetic Graph

In Section 3.3 we explained the different visit-degree ratios using an intuitive model, which comprised of two tightly connected regions with a sparse bottleneck connecting the two. In this section we generate synthetic graphs to examine the effect of regional graph properties on visit-degree ratio.

In order to reproduce the visit-degree phenomenon, we generate a graph with two components with a different average node degree. We set the degrees to span across the same range. We make two connected components each with $50,000$ nodes. Node degrees are drawn from ranges $[3, 5)$, $[5, 25)$, $[25, 125)$ and $[125, 625)$ with different probabilities. We can change average degree of each component by changing the probability of selecting nodes from each range. Two components used in this section have average degrees equal to 64.8 and 30.4.

Then we bridge components with different bottleneck widths. In order to bind components together, we rewire same number of links from both components. Figure 4 shows visit vs. degree plots of these graphs for different width bottleneck and length of random walk. Figures 4(b) and 4(c) compare the visit vs. degree plots of running $20,000$ random walks each with length 200 steps on two graphs, one with bottleneck twice as wide as the other one. Similar to 4.1 we observe that increasing the width of the bottleneck decreases the gap in visit-degree ratio of different components.

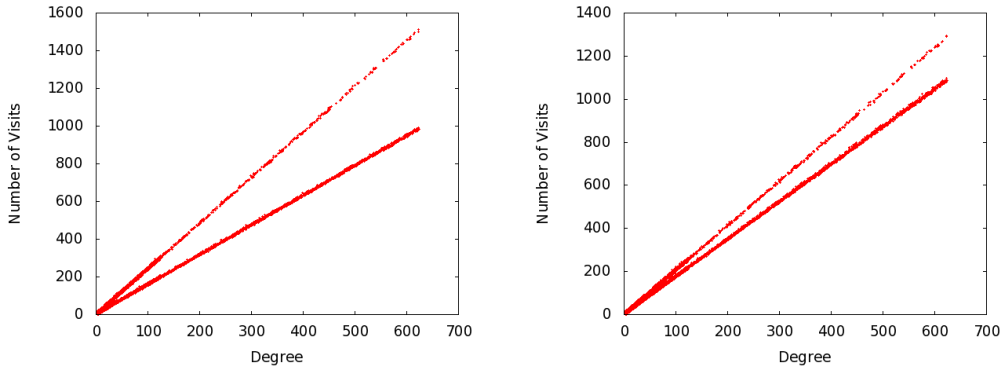
Comparing Figures 4(b) and 4(d) we observe the effect of random walk length. Random walks of latter figure are twice as long the former, over the same graph. Higher length of random walks increases their chance of crossing the bottleneck, hence alignment angels are brought closer.

In summary, we can state that lack of global proportionality of visit-degree can be traced back to different average degree of the regions of a graph.

4.3 Validation with High Degree Nodes Clusters

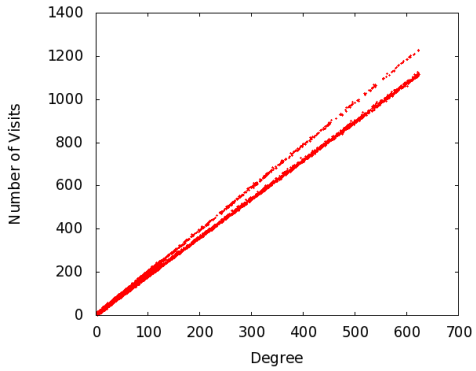
Thus far, our experiments suggest that visit vs. degree plots of short walks reveal regions of a graph. Nodes of each region line up along different angles based on average degree of the region's nodes. In this section we focus on a long random walk in its steady state to further investigate correctness of our hypothethis in snapshots of online social networks. To this end, we use snapshots of Orkut, LiveJournal, Flickr and YouTube which will be introduced in Section 6.

As shown in figure 1(a) in steady state visit-degree of all nodes are proportional for LiveJournal. We observed similar visit-degree proportionality for all other graphs that we use in this section. We focus on 500 highest degree nodes of each of the four graphs. Eliminating all other nodes from the walk trace, we calculate transition probability of node x to node y, where x and y are both in the set of highest degree nodes. This

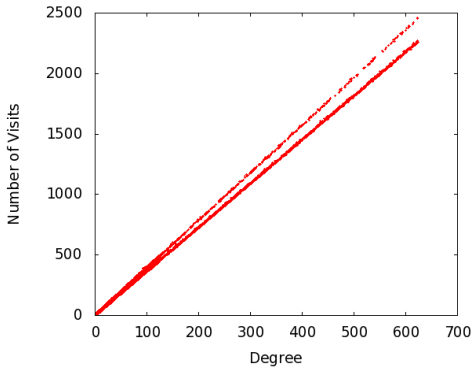


(a) Bottleneck=5K; Walk length=200

(b) Bottleneck=20K; Walk length=200



(c) Bottleneck=40K; Walk length=200



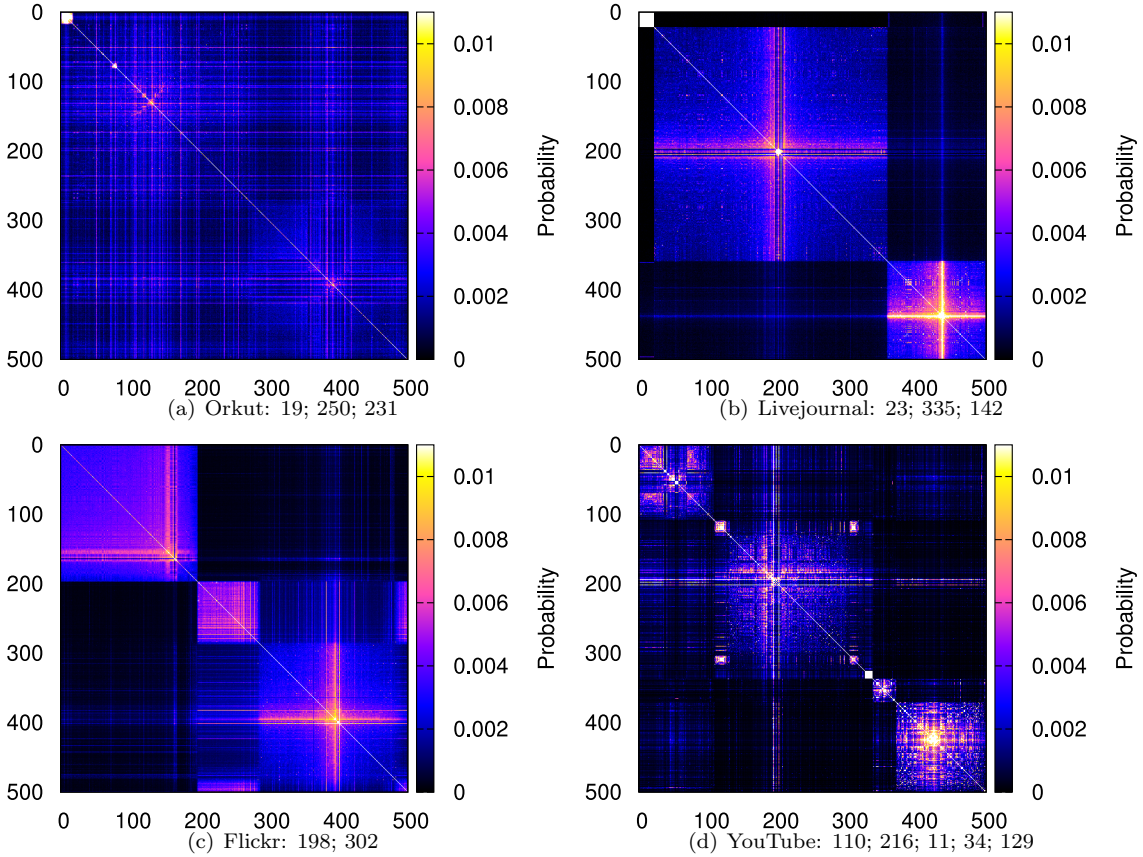
(d) Bottleneck=20K; Walk length=400

Figure 4: Synthetic graphs with two components and various width bottlenecks visits count vs. degree plots for different length of walkers

probability is the likelihood of visiting y immediately after x in the the reduced walk trace. Since random walks are inclined to stay in a region, we should see higher transition probabilities among high degree nodes of each region. We applied a spectral clustering technique [12] to the weighted full mesh of 500 highest degree node, where weight of a link is the transition probability of nodes on ends of the link. Figure 5 shows results from clustering these 500 nodes and presenting transition probabilities in the form of heat-maps. Nodes detected in one cluster are positioned side by side. Bright blocks in heat-maps show high likelihood of transition among nodes in each cluster which can represent regions of high degree nodes of graph. Size of each cluster is encoded in label of each figure, e.g. Orkut has 3 clusters of high degree nodes with 9, 250 and 231 nodes.

Ordering used in figure 5 is resulted of the clustering algorithm on transition probabilities. Here, we use visit-degree ratio of short random walks to line up nodes along axis of the heat-map. Similar to figure 2, short random walks on other three graphs also exhibit similar visit-degree disproportionality. We believe, since each local proportionality of visit-degree represent a region of the graph, this arrangement should also arrange nodes belonging to a region close to other nodes from the same region. Thus again bright blocks should appear in heat-maps using this ordering scheme. Figure 6 shows heat-map plots when nodes are ordered by visit-degree ratio. Bright blocks in transition probability heat-maps shows ordering by visits-degree ratio can line up a node that belong to a tightly connected region of a graph next to other nodes in the same region.

Comparing number and size of bright blocks in Figures 5 and 6 shows the two ordering scheme have similar



(a) Orkut: 19; 250; 231

(b) Livejournal: 23; 335; 142

(c) Flickr: 198; 302

(d) YouTube: 110; 216; 11; 34; 129

Figure 5: Transition probabilities among 500 highest degree nodes in each graph. Clusters of the graph are detected using the spectral clustering technique and nodes of each cluster are placed side by side along axis of the heat map

results. The main difference is that the first ordering is the result of a clustering algorithm which groups nodes of tightly connected regions, while the second one only arranges nodes based on nodes' visit-degree ratio. Although visit-degree ratio ordering does not result separated groups tightly connected clusters, it can separate them in its outcome ordering which are visually detectable in its heat-maps. It is only in Orkut graph that the spectral clustering algorithm splits the large not-so-bright block in top left corner of Figure 6(a) into two blocks in 5(a).

It also worth mentioning that separation of blocks are more clear for some graphs and vague in case of some other graphs. This is due to the fact that regions of different social graphs exhibit different levels of separation. For instance we expect YouTube to have less pronounced regions than other graphs since patterns in its heat map are more vague.

5 WalkAbout - Methodology

In this section we present our methodology called WalkAbout, for connectivity analysis of large graphs. While we apply this technique to extract connectivity features of graphs representing online social networks in the rest of the document, we believe this methodology is applicable to any other graph that exhibit regional connectivity features.

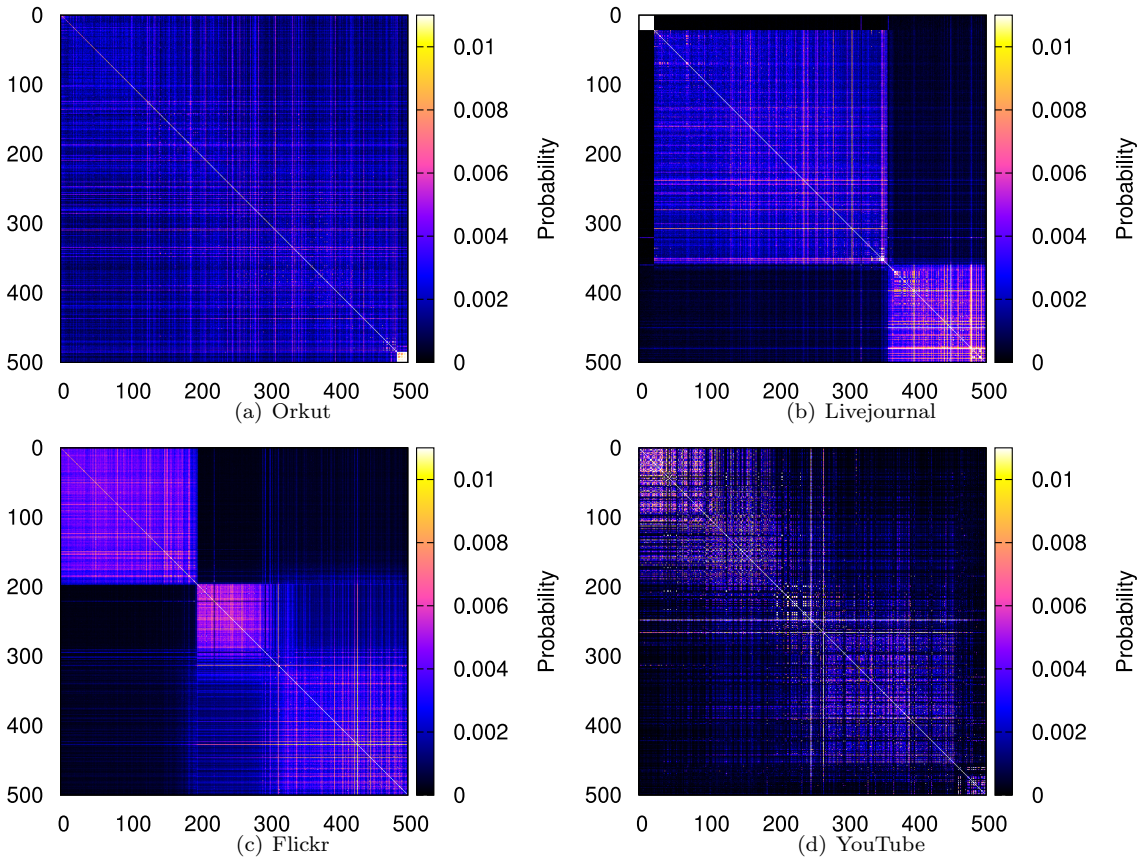


Figure 6: Transition probabilities among 500 highest degree nodes in each graph. Nodes are ordered by visit-degree ratio calculated by short random walks

The main ingredient of the WalkAbout technique is visit-degree ratio of nodes derived from many short random walk. Our main goal is to characterize regional connectivity of graphs, i.e. identify regions and their intra and inter connectivity.

For this purpose we need to:

1) Explore the transient random walk phase to discover a *critical* walk length. Increase in walk length, reduces the gap between visit-degree ratio angels, until only a single angel is visible in visit-degree plots. We call minimum walk length for which the visit-degree disproportionality disappears the critical walk length.

1) Cluster nodes based on their visit-degree ratio for random walks shorter than critical length. This basically aims to map nodes with different visit-degree ratios to different regions.

3) Derive global connectivity of the graph in terms of its regions detected as the result of previous step.

In the rest of this section, we briefly explain details of each of the above steps of WalkAbout. More detailed descriptions are provided in Section 7 and Section 8 as we apply WalkAbout on four snapshots of on-line social networks.

5.1 Sampling Graphs with Many Short Walks

WalkAbout entails performing a large number of short random walks, each one starting from a uniformly random starting node. Main parameters of this experiment are *a*) N: The number of walkers and *b*) L: The length of the random walk. In section 4.2, we briefly explained the effect of the walk length on the visit-degree ratio of nodes. Here we describe the effect of the number of walkers, as we dig deeper into the effect the walk length and introduce dynamics of regions with the increase of the walk length.

5.1.1 The Effect of Number of Walkers

In order to explain the effect of N we should recall its effect in Figures 1(b) and 1(c). While a single short walk is does not show visits count and degree proportionality because of the noise due to the small number of credits it can distribute, increasing the number of walkers increases the number of samples collected for estimation of the visit-degree ratio for a node, thus the proportionality patterns get more clear with increase in N. Thus as N grows, nodes move closer to the visit-ratio angel of their region. As a result, increasing N decreases variation of the visit-degree ratio. In case L is bellow the critical length, increasing N more clearly reveals separation of the visit-degree ratio of different regions.

We empirically examined the effect different number of random walks and fixed the number of walkers through out this phase analysis. The number of random walks performed on each graph equals to one fifth of the number of nodes in the graph.

5.1.2 The Effect of Walk Length

Local properties revealed by visit-degree ratio completely depend on percentage of walkers that do not cross graph's bottlenecks. Increasing L increases chances of random walks to cross these bottlenecks, thus reducing local properties displayed by walkers. As a result, Increasing length of random walk shifts properties portrayed by random walks towards those of the whole graph. Thus L has direct impact on separation of visit-degree ratios.

Since WalkAbout aims at capturing transient properties of graph, different length of random walks are examined to find critical random walk lengths. We gradually increase length of random walks and explore plots of visit vs. degree to check if we encounter clear gaps in visit-degree ratio. Steps performed by WalkAbout at this stage can be summarized as follows:

1. Start random walks from 20% of nodes each with length l.
2. Examine visit vs. degree scatter plots to see if critical length is reached, by looking at the number of clusters in the plot of visits vs. degree.
3. If critical walk length is reached ı.e. only a single alignment line exists in plots, then abort
4. Continue random walks for another l steps and goto step 2.

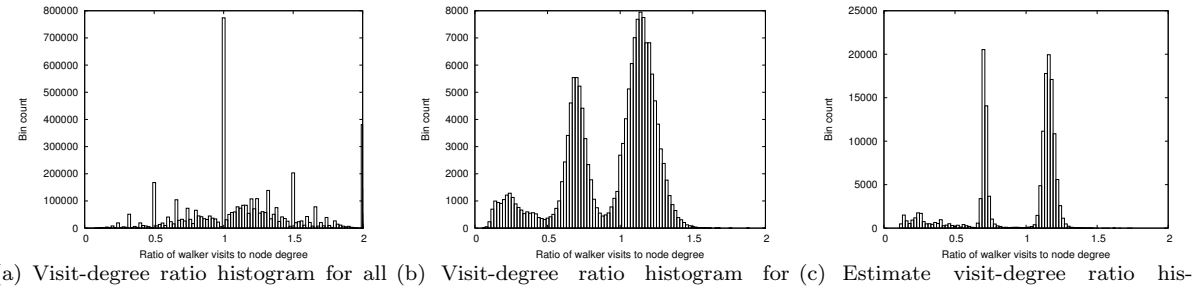
5.2 Clustering Based on Visit-Degree

In the previous section WalkAbout found the critical walk length. At this stage it clusters nodes with similar visit-degree ratio, to identify regions. We use a simple heuristic histogram clustering to detect nodes with similar visit-degree ratio. Figure 7 shows histograms of visit-degree ratios resulted from random walks with length 100 steps on LiveJournal.

On major challenge in histogram clustering is dealing with the noise in histogram imposed by low degree nodes. Visit count is inherently low for these nodes since each visit to the node drastically affects their visit-degree ratio. Thus, line up of the nodes along visit-degree ratio of their region is less reliable than high degree nodes. We can reduce variance of visit-degree ratio by increasing number of walkers, which increases cost of WalkAbout. Since the majority of nodes in the graph are low degree nodes, visit-degree histogram for all nodes visited by random walk would look like Figure 7(a). In order to reduce the noise of low degree nodes, we use visit-degree ratio histograms of high degree nodes. Figure 7(b) shows same histogram, when low degree nodes are not considered. Clearly new histogram is less noisy and dominant visit-degree ratios are more pronounced. WalkAbout examines different bin widths as well as different minimum degree thresholds for nodes contributing to the histogram to produce smooth histograms with clear minima and maxima.

Heuristic histogram clustering simply looks for local minima in histogram. Since histograms are still noisy we need to to define additional intuitive parameters which describe a valid local minima. These properties include depth of the minima compared to closest maxima and bin distances between two minima. Value of these parameters are graph specific and we tune the histogram clustering for each graph. In summary, histogram clustering for a graph should be personalized by setting i) minimum degree threshold, ii) bin width, iii) minimum acceptable depth of a minimum and iv) bin distance between two minima. Once tuned for a specific graph, histogram clustering has the complexity of $O(N)$ where N is number of histogram bins, thus its effectively scalable.

Outcome of clustering algorithm is ranges of visit-degree ratio. Each range spans around a maxima in the histogram and is used to separate nodes based on their visit-degree ratio.



(a) Visit-degree ratio histogram for all nodes

(b) Visit-degree ratio histogram for node degree > 100

(c) Estimate visit-degree ratio histogram for node degree > 100

Figure 7: 7(a): Visit-degree ratio histograms. 7(b): Visit-degree ratio histograms for nodes with degree higher than 100. Noise introduce by low degree nodes is removed in comparison to previous figure. 7(c): estimate visit-degree ratio histograms for nodes with degree above 100. Using estimated visit count clusters can be more accurately detected in comparison to previous figure

5.2.1 Using Estimated Visit Count

As mentioned in Section 5.1.1 increasing number of walkers increases the variance of visit-degree ratio. Considering aggregated visit count, its variance can be calculated using variance of binomial trial leading to the estimate visit count. Thus, number of trial is total length of all random walks and number of victories is total number of visits.

In this section we present the notion of estimated visit which has lower variance than aggregated visit count. WalkAbout processes random walk traces to evaluate estimated visit count. This value estimates number of times a node might have been visited given traces of all walkers. Each time visited, a node earns *one* credit when measuring aggregate visit count. On the other hand each neighbor of a node n with degree D earns $\frac{1}{D}$ estimated visit count credit, if n is visited, since they might be visited at the next step. Measuring estimated

visits count using latter method reduces the variance of visit count measurement, without imposing overhead of running extra random walks [3].

Figure 7(c) shows histogram of estimated visit-degree ratios for the same setting as Figure 7(b). Low variance of estimated visit count makes humps of the histogram more narrow while visit-degree histogram have wider humps. In the rest of this document we refer to estimated visit-degree ratio as visits-degree ratio and we use it for all histogram clusterings.

5.3 Regional Connectivity Evaluation

WalkAbout groups nodes that are likely to form tightly connected regions. WalkAbout then measures Intra and inter connectivity of regions in a graph to reveal its regional connectivity features. These features can be used to give a coarse summary of graph, by showing how its tight regions are connected together.

To this end WalkAbout evaluates *regional transition probability*, *regional transition odd* and *random walk distance* of region i to region j. WalkAbout also measures *random walk diameter* of a region by averaging number of consecutive steps a random walk stays in a region. More details about these measurements are provided in Section 8.

6 WalkAbout - Application and Evaluation

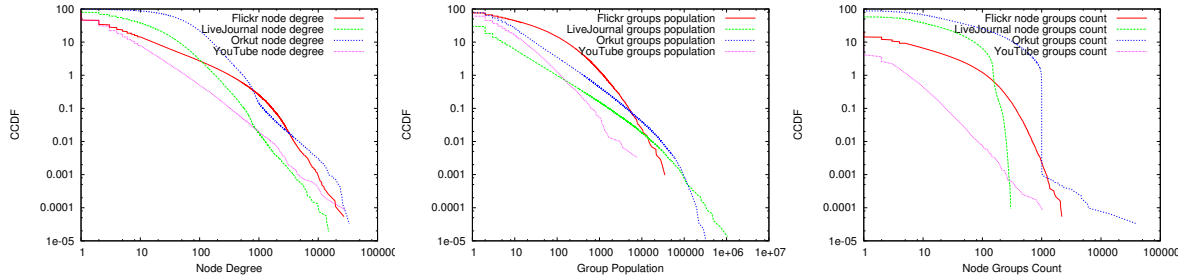
We obtain snapshots of four on-line social networks, namely Orkut, LiveJournal, Flickr and YouTube, provided by Mislove et al. [13] in order to demonstrate WalkAbout capabilities. In addition to the underlying social connectivity structure, these snapshots include group membership information. Groups are mainly used to gather users sharing a common interest and facilitate their communications. They enable users to share content and conversation, either privately or publicly. User can join or be joined to a group by its owner based on its moderation type and openness. High-level statistics of these snapshots are provided in Table 1. Distribution of node degree, groups population and number of groups a user has joined are shown in Figure 8. Since WalkAbout is designed to detect regions of a single connected component, it only uses the main connected component for its analysis.

Table 1: High-level statistics of studied on-line social networks

	Orkut	LiveJournal	Flickr	Youtube
Number of nodes	3,072,627	5,203,764	1,861,233	1,157,828
Number of singleton nodes	186	80,694	145,978	19,329
Size of largest connected component	3,072,441	5,189,809	1,624,992	1,134,890
Number of edges	234,370,166	97,419,546	31,110,082	5,980,886
Number of groups	8,730,860	7,489,297	103,649	30,088
Crawl Date	Oct 3 - Nov 11, 2006	Dec 9 - 11, 2006	Jan 9, 2007	Jan 15, 2007

Process of application of WalkAbout to these graphs and measurements of detected regions comprise of following tasks:

[3] Lower variance of estimated visit count is due to smoother variation of estimated visit count in comparison to aggregate visit count, as random walks are performed. Visit count increases with values of 1 as walkers walk the graph, but variations in estimated visit count are smaller at each step.



(a) Distribution of nodes' degree in each graph (b) Groups population distribution (c) Distribution of number of groups users participate

Figure 8: Properties of graphs and groups in the graph

1) We examine evolution of the population and mean visit-degree ratio of detected regions as we increase random walk length. We also demonstrate how increasing the length of the random walks results in merging of close regions.

2) We show a regional summary of a graph. This summary contains: *i*) The size of each region, *ii*) regional transition probability, *iii*) regional transition odd, *iv*) random walk distance between regions and *v*) average number of steps a random walk stays in a region namely random walk diameter of a region

3) We evaluate internal and external connectivity of regions detected by WalkAbout by counting the number of links that connect nodes of a region together and number of links that connect a region to other regions. We also compare these numbers with probabilistic evaluation of corresponding random graph.

4) Finally we investigate root causes leading to formation of the detected regions. We use group membership information which were provided along the snapshots we obtained.

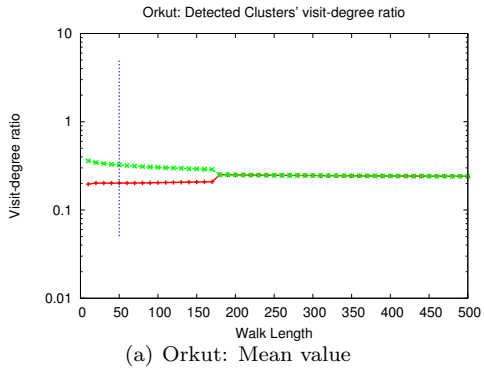
7 Evolution of Region Characteristics with Increase in Walk Length

As we mentioned earlier, increase in the length of random walks results in increase in the number of random walks reaching regions other than the region of their starting point. Increase in number of walkers crossing bottlenecks reduces separation of visit-degree ratio of nodes of different regions. This makes detection of separate regions impossible when walk length exceeds a certain value. At this point we consider two regions merged since WalkAbout can not detect them. *Critical walk length* is thus introduced as a random walk length beyond which no separations in visit-degree ratios are discernible.

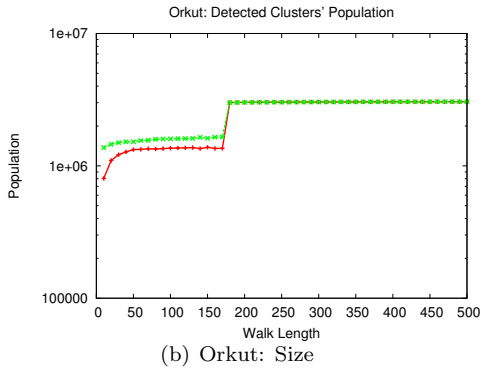
Dynamics of regions detected by different walk length disclose interesting information about connectivity of regions. Intuitively regions with wider bottleneck should merge in shorter walk lengths. In this section we measure dynamics of regions detected by our heuristic histogram clustering by evaluating sizes of regions and their average visit-degree ratio. As a result we can visualize region sizes and average visit-degree ratio, as well as walk length in which regions merge.

For each OSN graph, we perform random walks and apply the region detection technique based on visit-degree ratio at each 10 steps, until random walks reach the length of 500 steps. This way we explore dynamics of regions to locate critical walk length. In order to map regions of walk length t to regions of random walk length t^+ we keep track of nodes' regions at walk length t and t^+. We call region x of walk length t the same as region x' of walk length t^+ if majority of nodes in region x end up being members of region x'.

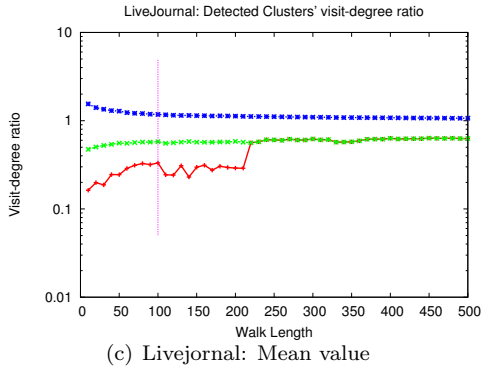
Figures 9(a) through 9(h) present regions' dynamics by illustrating evolutions of regions' sizes and the mean value of visit-degree ratio as we increase the length of random walks. Figures show that number of regions



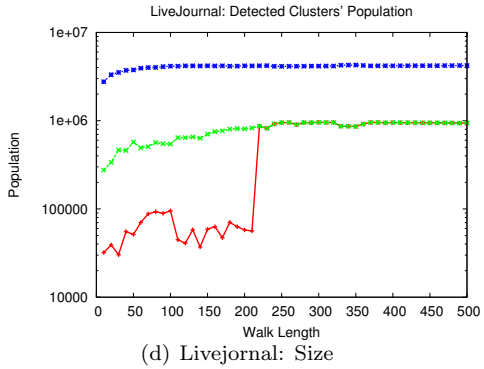
(a) Orkut: Mean value



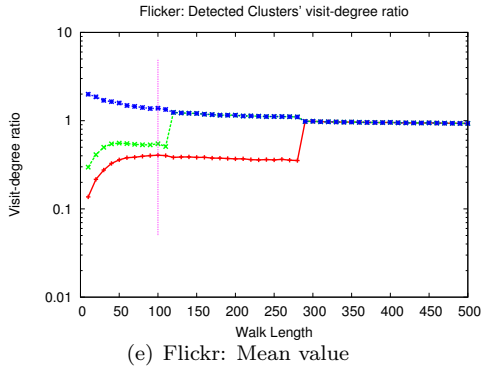
(b) Orkut: Size



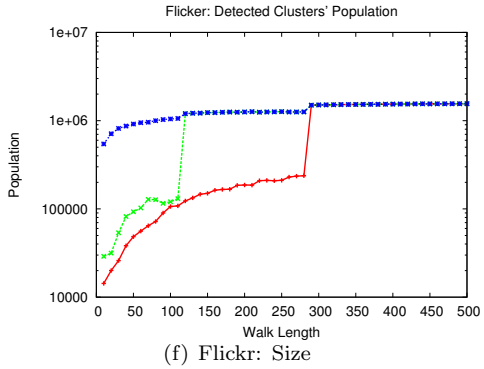
(c) Livejornal: Mean value



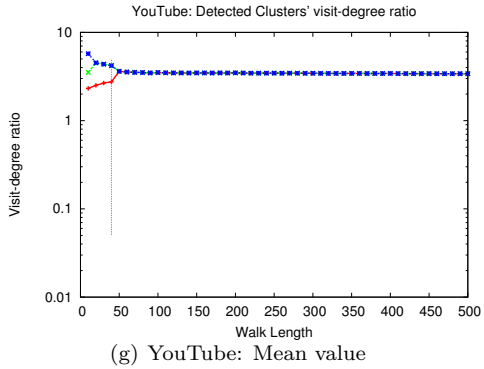
(d) Livejornal: Size



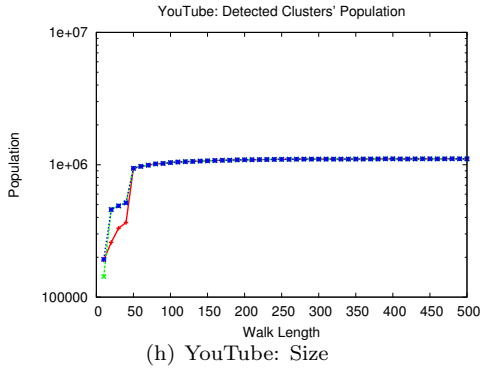
(e) Flickr: Mean value



(f) Flickr: Size



(g) YouTube: Mean value



(h) YouTube: Size

Figure 9: Evolution of visit-degree ratio and size of clusters detected in different social networks

detected for each OSN is different. While two regions were detected in Orkut and YouTube, the number of detected regions is three for LiveJournal and Flickr before critical length of random walk is reached. We assign region identifiers by their mean visit-degree ratio. Thus $R0$ has lowest visit-degree ratio and appears below all other ones in graphs representing evolution of visit-degree ratio.

Evolution of visit-degree ratio suggests as we increase the length of random walks the gap between visit-degree ratio of region decreases. Similar explanation as the one given for synthetic graphs in Section 4.2 can be made here. Growth in walk length increases number of walkers which cross graphs bottlenecks, hence closes the gap between regions visit-degree ratios. Evolution of sizes of regions suggests as we increase length of random walkers, more nodes are visited hence we can assign them to their corresponding region. These plots suggest Orkut is comprised of two regions with similar sizes while region sizes in other OSNs are more diverse. As we mentioned earlier increasing the length of random walk merges regions form the perspective of WalkAbout. Two lines joining represent a merge in these plots. Intuitively regions that merge in shorter walk lengths are closer in the graph, i.e. bandwidth of bottleneck connecting the two is wider.

Based on these results that show dynamics of regions in the graph, we select a *proper length* for random walk and use the regions of the chosen random walk length to study static regional view of the graph. We select proper random walk length in such a way that maximum number of region is detected before walks pass their critical length. In addition, since increase in length of walkers linearly increases time complexity of WalkAbout, a stable exploration point of shorter walk length is more desirable. In other words, we are interested in the largest possible distinct regions with minimum walker length.

We select walk lengths of 50, 100, 100 and 40 for Orkut, LiveJournal, Flickr and YouTube respectively. We consider all nodes that are not assigned to a region [4] as a part of an $Unknown$ region. This region also includes nodes that have not been visited by random walks. These regions are used in the next section, in which we evaluate regional connectivity of graph using random walks.

8 Graph Summarization - Region Representation

In this section, we extract connectivity of detected regions using random walk traces. We use five metrics, derived from random walks, to summarize connectivity features of each graph's regions. These metrics include, *i*) *random walk transition probability* among regions' nodes, *ii*) *random walk transition log odd* of transitions between all regions, *iii*) *random walk distances* of different regions and *iv*) *random walk diameter* of regions. We introduce each metric and show how each of them characterizes connectivity features of regions.

The goal of this section is to represent a zoomed out view of a graph. To this end we process random walk traces to produce a regional view of a random walk trace as in Figure 10. In this view of the random walk each node is marked by its region. Figure 10 shows node view of a trace and its corresponding region view. In this figure, nodes $x1$, $x2$, $x3$ and $x10$ belong to $R1$, $x5$, $x8$ and $x9$ are members of $R2$ and the rest of the nodes are $Unknown$. A transition in random walk from Ri to Rj occurs when a random walk visits node Rj after visiting Ri. $Unknown$ nodes can be visited during a transition. All transitions in this trace are shown in the processed view of the trace. As Figure 10 shows, $x4$ is $Unknown$ and was visited during a transition from $R1$ to $R2$. Thicker arrows in the processed view represent inter-region transitions, while thin arrows represent intra-region transitions. Transition probability of Ri to Rj is measured as the number transition from Ri to Rj divided by the total number of transitions from Ri.

$$TransProb_{i \to j} = \frac{|trans_{i \to j}|}{\sum_{k \in Regions} |trans_{i \to k}|} \tag{4}$$

[4] Since cluster bounds are selected using histogram of visit-degree ratio of nodes with degree higher than a threshold, it's possible to have nodes not mapped to a region.

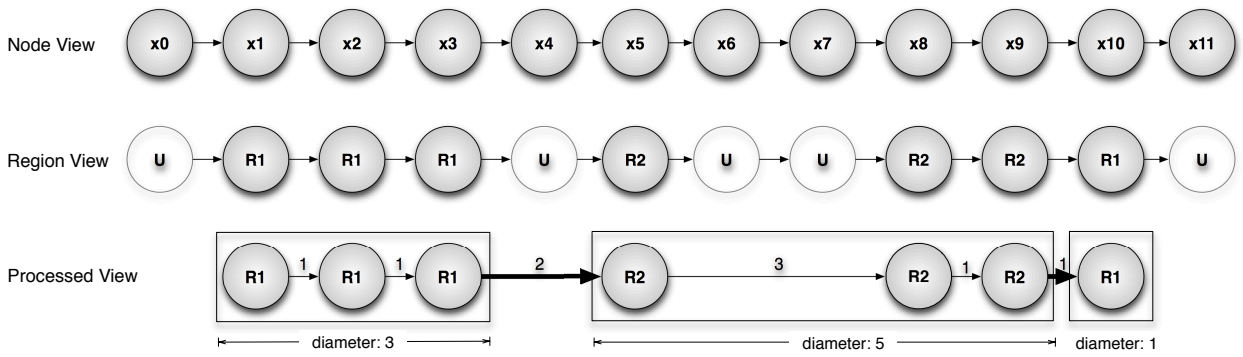


Figure 10: Schematic trace of a random walk

Transition Probability is biased toward regions with large number of nodes. In other words, the larger the number of nodes in a region, the higher the chance of visiting that region in future steps of random walk even if detected regions were completely random. Transition odd on the other hand measures how unlikely a transition event is, given the regions' populations. Equation 5 measures oddness of transition from Ri to Rj. Here P_j fraction of the population is assigned to region Ri. Positive value $TransOdd_{i\to j}$ shows higher chance of visiting a node from Rj in next steps, when the random walk is currently visiting Ri.

$$TransOdd_{i\to j} = \log\left(TransProb_{i\to j}\right) - \log\left(P_j\right) \tag{5}$$

Number of $Unknown$ nodes visited during a transition exposes transition distance of two regions. Transition distance can also measure number of $Unknown$ nodes visited between consequent visits of nodes from the same region. This evaluates penetration of $Unknown$ nodes among nodes of a region. In Figure 10, the weight of each arrow shows random walk distance of regions involved in that transition. In order to evaluate random walk distance from Ri to Rj, transition distances of all transitions from Ri to Rj are averaged.

Tightly connected regions confine random walks for a large number of steps. Sparse connectivity of a region to other regions prevents a random walk from escaping the region, thus a random walk stays in a region for a large number of steps. Random walk diameter measures average number of steps, a random walk stays in a region. Duration of visit to Ri starts and finishes by visits to nodes from the Ri, while nodes in between two visits can be a combination of nodes from the Ri and $Unknown$. Boxes in processed view of Figure 10 represent *three* visits to regions in this trace. Figure shows *two* visits with duration *three* and *one* to $R1$ and *one* visit with duration $five$ to $R2$. Random walk diameter of a region is its average duration of visits. Random walk diameter shows average number of steps a random walk takes till visiting a node from other regions.

Size of regions, regional transition probabilities, regional transition odds, random walk distances and random walk diameters of regions are used to measure connectivity feature of graph using regions detected in the proper walk length.

Tables 2 through 5 summarize regional connectivity features of the studied OSNs. Evaluations were performed for two clustering scenarios: 1) The entire detected cluster is considered as a region. 2) 10% highest degree nodes of each cluster belong to region and rest are in the $Unknown$ region. Results are provided in "Whole region" and "High Deg" columns of each table respectively. All regional summary metrics are evaluated using a single random walk in steady state, otherwise transitions are biased toward short distance transitions due

Table 2: Orkut: Graph summarization indices

	Whole region								High Deg	
	Size	Trans Prob		Trans Odds		RW Dist		Diam	RW Dist	Diam
	R_i	$R0$	$R1$	$R0$	$R1$	$R0$	$R1$	R_i	R_i	R_i
$R0$	43.2%	**99.26%**	0.73%	0.32	−1.8	1.00	1.02	135	2.63	322
$R1$	49.4%	1.10%	**98.91%**	−1.6	0.26	1.02	1.00	90	7.65	207
U	7.3%	35.78%	64.21%							

Table 3: LiveJournal: Graph summarization indices

	Whole region											High Deg	
	Size	Trans Prob			Trans Odds			RW Dist			Diam	RW Dist	Diam
	R_i	$R0$	$R1$	$R2$	$R0$	$R1$	$R2$	$R0$	$R1$	$R2$	R_i	R_i	R_i
$R0$	2.1%	**95.37%**	3.90%	0.72%	1.61	−0.57	−2.06	1.00	1.02	1.07	21	2.38	221
$R1$	13.3%	1.10%	**93.15%**	5.73%	−0.32	0.80	−1.16	1.01	1.00	1.00	14	2.39	107
$R2$	75.3%	0.05%	1.62%	**98.32%**	−1.60	−0.96	0.07	1.07	1.00	1.00	59	2.47	404
U	9.2%	6.52%	27.32%	66.14%									

to short length of the random walks [5]. Transition probabilities from $Unknown$ region to Ri on are estimated by measuring the percentage of short walkers starting from an $Unknown$ node which visit a node in Ri before visiting nodes from all other regions. The rest of the walker is simply ignored after the jump from $Unknown$ to Ri.

In the following, we briefly point out details of these summaries:

First, transition probability and transition odd, show high likelihood of visiting a node in a region after a node from same region is visited. This simply shows that random walks get stuck in regions. Transition probabilities suggest regions are more internally connected.

Second, each graph exhibits different levels of regional separation. Based on transition probabilities the highest level of regional separation is exhibited in Orkut. LiveJournal, Flickr and finally YouTube show lower regional separation. Regional diameter evaluated for whole regions and high degree regions also support the same ordering of regional separation.

Third, transition odd divulges reachability information among regions. Regions with higher average transition odds ($AvgTransOdd_{i,j} = \frac{TransOdd_{i\to j}+TransOdd_{j\to i}}{2}$) merge in shorter walk length. Figure 9(e) shows that first regions who merge in Flickr are $R2$ and $R3$. Thus we argue that the bottleneck connecting $R2$ and $R3$ should be wider than bottleneck of $R1$ and $R2$. This agrees with the fact that $AvgTransOdd_{1,2}$ is -0.655 while $AvgTransOdd_{2,3}$ is -0.475. On the other hand in case of LiveJournal $AvgTransOdd_{1,2} = -0.73$

[5] Short random walks do not see longs distance transitions and duration of visits

Table 4: Flickr: Graph summarization indices

	Whole region											High Deg	
	Size	Trans Prob			Trans Odds			RW Dist			Diam	RW Dist	Diam
	R_i	$R0$	$R1$	$R2$	$R0$	$R1$	$R2$	$R0$	$R1$	$R2$	R_i	R_i	R_i
$R0$	5.3%	**94.84%**	1.77%	3.38%	1.08	-0.75	-1.39	1.00	1.01	1.02	18	1.36	31
$R1$	6.7%	1.87%	**82.29%**	15.82%	−0.62	0.92	-0.72	1.01	1.00	1.00	5	1.47	7
$R2$	56.0%	1.32%	5.85%	**92.82%**	−0.77	-0.23	0.05	1.02	1.00	1.00	14	1.35	24
U	31.8%	14.11%	8.78%	77.10%									

Table 5: YouTube: Graph summarization indices

	Whole region								High Deg	
	Size	Trans Prob		Trans Odds		RW Dist		Diam	RW Dist	Diam
	R_i	$R0$	$R1$	$R0$	$R1$	$R0$	$R1$	R_i	R_i	R_i
$R0$	30%	**78.32%**	21.67%	0.30	−0.45	1.005	1.009	4	1.39	8
R	47.5%	21.53%	**78.46%**	−0.25	0.10	1.009	1.021	4	2.07	11
U	22.3%	44.58%	55.41%							

and $AvgTransOdd_{2,3} = -1.06$ suggest merge of $R1$ and $R2$ happens in a shorter walk length which is also supported by Figure 9(c).

Forth, since intra-region distances are lower than inter-region distances, we conclude that *Unknown* nodes are mostly spread in between regions rather than inside a region. Fifth, spread of low degree nodes of a region between its high degree nodes can be inspected by the changes in intra-region distances, for whole region and its 10% high degree nodes. This increase is approximately 100% for all regions in all graphs, except for Orkut's $R2$, for which 600% increase in random walk distance. In case of Orkut, this change suggests that high degree nodes in each region are mostly connected through lower degree nodes in that region.

It is worth mentioning that random walk diameter of regions increases for regions with only high degree nodes although we have reduced the size of regions to 10% of its original size. High variance visit-degree ratio for low degree nodes reduces the confidence by which we can estimate their visit-degree ratio. This decreases the confidence of cluster detection for them. Thinning out regions by simply focusing on their higher degree nodes is equivalent to clustering nodes with high visit-degree ratio variance to the *Unknown* region.

We conclude that WalkAbout is capable of finding tightly connected regions and characterizing graphs based on connectivity of detected regions.

8.1 Regional OSN Representation

Connectivity of OSNs can be characterized as their regional connectivity features. Figure 11 simply represents a schematic representation of OSNs that we studied. Figures present region metrics evaluated in Section 8. Each region is represented with a circle. The size of each circle is proportional to the proportional size of the region it represents. The color of each circle represent region's relative diameter to its size. The darker the color, the higher the capability of region in restraining random walks from escape. Transition probabilities are coded in transparency of arrows connecting regions. The more opaque the arrow, the higher the transition probability. In case a graph is comprised of more than two regions, circles that represent closer regions - those that have larger average random walk transition odd - are placed closer in the schematic figure.

We can evidently see from these figures that, 1) Orkut and YouTube have two region with comparable sizes. Flickr and LiveJournal on the hand are comprised of three region with different sizes. 2) intra-region transitions are more likely in all graphs among all graphs transitions. Comparing intra region transitions of YouTube with others suggests higher randomness. Even-though its not clear due to limitation of these plots, inter region transitions of YouTube are twice as much probable than higher intra region transitions in other OSNs. 3) normalized region diameter shows higher internal connectivity in Orkut and small regions Flickr and LiveJournal.

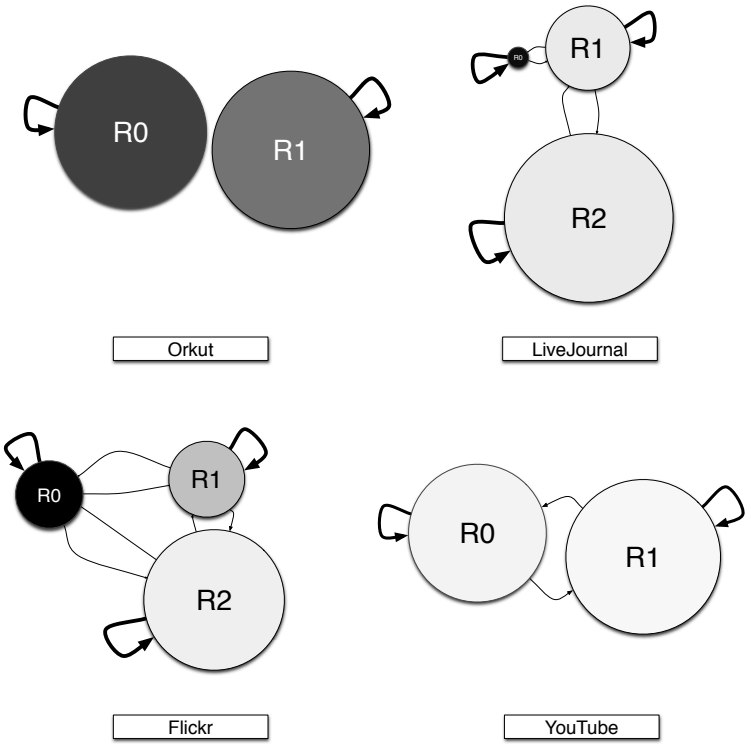


Figure 11: Regional OSN representation

9 Graph Investigation

In this section, we map node level connectivity to regions of the graph to evaluate exact connectivity of detected regions. We label each node by the its region. Then we count the number of links internal to a region and the number of links that connect two regions. Percentages of internal and external links are provided in Tables 6 through 9 for different OSNs. In each table, $C_{i\to j}$ is portion of Ri's links which connect Ri to Rj. These results show that intra-region links account for majority of region's connections, which aggress with high level of intra-region connectivity deducted from graph summarization.

Table 6: Orkut: edge division among regions and region size

Edges: $C_{i\to j}$				Nodes		$R_{*\to j}$	Connectivity		
	$R0$	$R1$	U	Size	m_deg	e_links	CC_size	CC_m_deg	CC'_m_deg
$R0$	**97.32%**	1.41%	1.28%	43.2%	105.44	59.76%	98%	106.8	9.2
$R1$	2.12%	**96.37%**	1.52%	49.4%	60.98	39.53%	99%	61.2	12.7
U	55.13%	43.56%	1.32%	7.3%	7.39	0.71%			

$C_{i\to j}$ values can be represented in the form of a matrix. Matrix C can be used to calculate traditional cluster evaluation indices such as *internal conductance*. Conductance is an index founded on the concept of bottlenecks, i.e. a cluster should not have a sparse cut separating non-trivial parts and in contrast, the connection of a cluster to the remaining graph should be sparse [14]. Thus, internal conductance of Ri equals $(1 - C_{i\to i})$, which reveals its sparse connection to the rest of the graph for all detected regions.

Table 7: LiveJournal: edge division among regions and region size

Edges: $C_{i\to j}$					Nodes		$R_{*\to j}$	Connectivity		
	$R0$	$R1$	$R2$	U	Size	m_deg	e_links	CC_size	CC_m_deg	CC'_m_deg
$R0$	**90.33%**	7.21%	1.29%	1.17%	2.1%	50.98	5.81%	33%	132.4	12.2
$R1$	1.99%	**86.19%**	10.57%	1.26%	13.3%	28.564	20.66%	60%	44.1	5.0
$R2$	0.11%	3.15%	**95.91%**	0.83%	75.3%	17.86	72.96%4	99%	17.96	2.7
U	6.75%	26.37%	58.56%	8.33%	9.2%	1.36	0.57%			

Table 8: Flickr: edge division among regions and region size

Edges: $C_{i\to j}$					Nodes		$R_{*\to j}$	Connectivity		
	$R0$	$R1$	$R2$	U	Size	m_deg	e_links	CC_size	CC_m_deg	CC'_m_deg
$R0$	**89.02%**	3.31%	6.27%	1.39%	5.3%	67.95	21.64%	28%	239.0	3.1
$R1$	3.14%	**69.41%**	26.64%	0.82%	6.7%	50.83	20.49%	45%	110.4	3.8
$R2$	2.39%	10.73%	**84.94%**	1.94%	56%	16.55	55.55%	98%	16.87	1.7
U	10.06%	6.25%	36.77%	46.91%	31.8%	1.60	2.33%			

We compute $R_{i\to j}$, a probabilistic metric, which is the probability of seeing a totally random node at the other end of links of Ri as we measure $C_{i\to j}$. $R_{i\to j}$ only depends on the number of nodes in Rj and average degree of nodes in Rj. In other words, it depends on the total degree of Rj. Equation 6 explains how $R_{i\to j}$ is estimated.

$$R_{i\to j} = \frac{|Rj| \times AvgDeg(j)}{\sum_{k\in regions} |Rk| \times AvgDeg(k)} \tag{6}$$

Comparison of $C_{i\to j}$ and $R_{i\to j}$ also reveals high level of internal connectivity. For example $R1$ in Orkut graph would have had 39.53% of its links as internal links in the random connectivity, but 96.37% of $R1$'s links are connected internally which shows high level of clustering for this region.

We also examine Internal connectivity of regions using flooding to check whether each region is a single connected component. We evaluate CC_size which is percentage of region's nodes that are part of the connected component. Although the majority of links in a region are internal links, none of these regions are a single connected component. Tables 6 through 9 also contain results of this inspection. In some cases, The largest Connected component is small in comparison to the region itself. WalkAbout may detect two regions with similar average degree as a single region. To check if detected regions are union of two clusters in the graph, we examine nodes that are not in the connected component. These nodes do not form another large connected component, i.e. their connectivity is in the form of many separate trees. As we compare average degree of nodes in the connected component (CC_m_deg) to nodes that are not in the connected component (CC'_m_deg), we can see the majority of nodes outside the connected component are low degree nodes. Average degree of nodes in the connected component is an order of magnitude higher than that of the rest of the region. Thus, we conclude that each detected region is union of a single connected component

Table 9: YouTube: edge division among regions and region size

Edges: $C_{i\to j}$				Nodes		$R_{*\to j}$	Connectivity		
	$R0$	$R1$	U	Size	m_deg	e_links	CC_size	CC_m_deg	CC'_m_deg
$R0$	**60.84%**	33.64%	5.52%	30.0%	8.07	47.00%	73%	10.2	2.07
$R1$	33.57%	**61.04%**	5.39%	47.5%	5.12	47.17%	93%	5.3	1.75
U	35.75%	35.00%	29.24%	22.3%	1.40	5.83%			

and many small low degree components.

This can be explained by the high variance of visits count estimation for low degree nodes. Variance of visit count makes assignment of low degree nodes to their correct regions using visit-degree ratio a difficult task. Multi partition structure of regions cannot be detected when connectivity is examined by percentage of internal links. This is due to the fact that sum degree of low degree nodes outside of the region's connected component is not comparable to sum of degrees of nodes in the connected component. Thus their connections to other nodes does not play an important role in evaluation of $C_{i \to j}$.

10 Investigating Root Causes

Thus far we applied WalkAbout to extract regions of tightly connected nodes in large scale social graphs. These regions have high level of internal connectivity and bottlenecks connecting regions are sparse. In this section we briefly explore why regions of tightly connected nodes form in a social graph. Intuitively users with similar interests and properties can form tightly connected clusters. Thus, investigating the correlation between user properties could reveal the reason behind formation of regions.

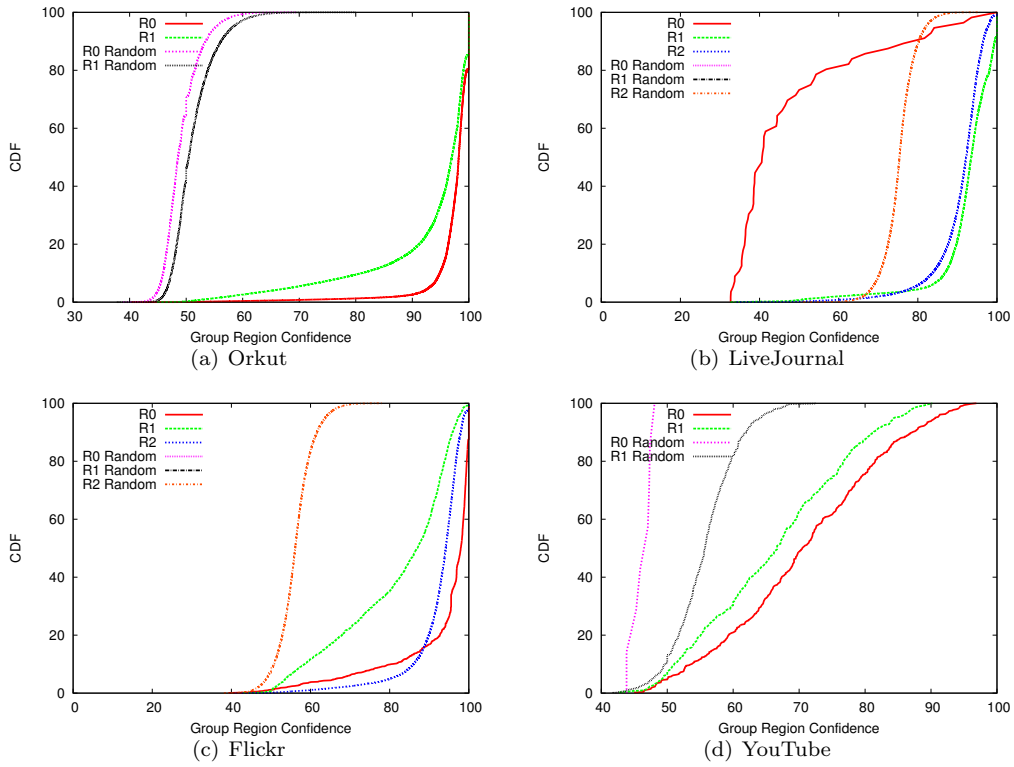
User and group identifiers are anonymized in our snapshots. Dealing with anonymized user identifiers, we have no information relating user identifiers to their geographical and social properties and their interests. Hence, we can not use user properties to see if a correlation exists among social properties of users in a region. We use group membership information collected alongside the crawl of each snapshot. Groups are implemented in many on-line social networks to allow users with similar interests to interact in a group in one-to-many fashion via posting messages and uploading content. In order to investigate and extract the underlying root causes for formation of these clusters we augment social graphs which with group memberships, to assemble an affiliation network [15]. In the augmented graph, links both connect users to their friends and groups [6] to their members.

Kleinberg, in [1] suggests that co-evolution of social and affiliation networks, results in clusters of tightly connected regions in social network, thus majority of members of a group should belong to a single region. In order to examine this hypothesis, we introduce *group region confidence* as an index to measure distribution of group members among regions. Group g is assigned to region R with confidence $X\%$, if majority of its members belong to R and the majority comprise $X\%$ of its total population. Since group size can be used as measure of group activity we only measure the index for groups with population above 50 to focus on groups revealing common interests of users.

We compare *group region confidence* with the case in which each member's region is selected randomly from a distribution proportional to each region's population. In the latter case, having region R containing 80% of users, each group member is assigned to R with probability 0.8. This gives us a good baseline for comparison, to show if groups are aligned with regions in the graph, i.e. if they are confined within boundaries of a region.

We plot the distribution of *group region confidence* for groups assigned to each region separately alongside that of the random case in Figure 12. From these plots it is clear that majority of a group's members are mapped to single region and *group region confidences* are much higher compared to the random case. It is worth mentioning that some regions are so small that no region will be assigned to them in the random case. In this case, we can not see a group region confidence CDF for the random case in 12. $R0$ and $R1$ in LiveJournal and Flickr are example of such regions. Interestingly in the real case, these regions still have groups assigned to them with high confidence as represented in Figures 12(b) and 12(c). Evaluation of *group region confidence* supports our initial hypothesis. We conclude that regions and groups are aligned and tight connected regions is the result of social and affiliation co-evolution of the network.

[6] Group represent a foci as is the terminology of affiliation networks.



(a) Orkut (b) LiveJournal

(c) Flickr (d) YouTube

Figure 12: Evaluation of membership of groups to regions

11 Conclusion and Future Work

WalkAbout characterizes connectivity feature of large graphs. To this end it presents a coarse view of a graph by detecting graph's tightly connected regions. It's very scalable since it merely uses visit counts collected by collection of short random walks and as long as random walks are applicable, it can be applied to any graph. Application of WalkAbout does not rely on availability of the full graph's snapshot since random walks can be performed online. It also can be parallelized by concurrently performing multiple random walks on the graph. Results from application of WalkAbout shows that WalkAbout is capable to detecting tightly connected regions in large graphs of online social networks. WalkAbout also summarizes graph's connectivity using it regional features. Its only shortcoming is in detecting low degree node's regions. Small number of visits to these nodes makes visit-degree ratio a high variance signal for these nodes. Thus confidence of region detection is low for these nodes.

As of future work, we believe detected regions can be further improved by taking into account information of individual random walk traces. Having high degree nodes of a region as region's anchors, we can introduce a level of confidence to region assignment for each node based on number of appearance of node on short random walks mapped onto a region. Hierarchical application of WalkAbout on the detected regions can help identify subregion in the graphs. WalkAbout can be applied to other large scale graphs with regional properties as well. WalkAbout can also be used to investigate evolution of the graph connectivity as users are added to the system and as social ties are introduce and removed by users.

References

[1] D Easley and J Kleinberg, "Networks, Crowds, and Markets," *Cambridge Univ Press*, pp. 85–118, 2010.

[2] David Harel and Yehuda Koren, "On clustering using random walks," *FST TCS 2001: Foundations of Software Technology and Theoretical Computer Science*, pp. 18–41, 2001.

[3] Haifeng Yu, Michael Kaminsky, Phillip B. Gibbons, and Abraham D. Flaxman, "SybilGuard: Defending Against Sybil Attacks via Social Networks," *IEEE/ACM Transactions on Networking*, vol. 16, no. 3, pp. 576–589, June 2008.

[4] Haifeng Yu, Phillip B. Gibbons, Michael Kaminsky, and Feng Xiao, "SybilLimit: A Near-Optimal Social Network Defense against Sybil Attacks," *2008 IEEE Symposium on Security and Privacy (sp 2008)*, , no. Figure 1, pp. 3–17, May 2008.

[5] G. Danezis and P. Mittal, "Sybilinfer: Detecting sybil nodes using social networks," in *Interner Bericht. Fakultät für Informatik, Universität Karlsruhe; 2006, 24*. 2009, NDSS.

[6] M Rosvall and C T Bergstrom, "Maps of random walks on complex networks reveal community structure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 4, pp. 1118–1123, 2008.

[7] M. Maggioni and R.R. Coifman, "Multiscale analysis of data sets with diffusion wavelets," .

[8] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 8, pp. 888 –905, aug 2000.

[9] Deepak Verma and Marina Meila, "A Comparison of Spectral Clustering Algorithms," Tech. Rep., Department of CSE University of Washington Seattle, WA 98195-2350, 2005.

[10] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. 2001, pp. 849–856, MIT Press.

[11] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul Erdos is Eighty*, vol. 2, no. 1, pp. 1–46, 1993.

[12] Aaron Clauset, M. E. J. Newman, and Cristopher Moore, "Finding community structure in very large networks," *Physical Review E*, vol. 70, no. 6, pp. 066111+, Dec. 2004.

[13] Alan Mislove, Massimiliano Marcon, Krishna P. Gummadi, Peter Druschel, and Bobby Bhattacharjee, "Measurement and analysis of online social networks," *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement - IMC '07*, p. 29, 2007.

[14] Daniel Delling, Marco Gaertler, R. Görke, Zoran Nikoloski, and Dorothea Wagner, "How to Evaluate Clustering Techniques," *Interner Bericht. Fakultät für Informatik, Universität Karlsruhe; 2006, 24*, , no. 001907, pp. 1–12, 2006.

[15] Ronald L. Breiger, "The Duality of Persons and Groups," *Social Forces*, vol. 53, no. 2, pp. 181–190, 1974.