Inferring Coarse Views of Connectivity in Very Large Graphs

Reza Motamedi  
University of Oregon  
motamedi@cs.uoregon.edu

Reza Rejaie  
University of Oregon  
reza@cs.uoregon.edu

Daniel Lowd  
University of Oregon  
lowd@cs.uoregon.edu

Walter Willinger  
Niksun, Inc.  
wwillinger@niksun.com

Roberto Gonzalez  
UC3M, Madrid  
grgonza1@it.uc3m.es

ABSTRACT
This paper presents a simple framework, called WalkAbout, to infer a coarse view of connectivity in very large graphs; that is, identify well-connected “regions” with different edge densities and determine the corresponding inter- and intra-region connectivity. We leverage the transient behavior of many short random walks (RW) on a large graph that is assumed to have regions of varying edge density but whose structure is otherwise unknown. The key idea is that as RWs approach the mixing time of a region, the ratio of the number of visits by all RWs to the degree for nodes in that region converges to a value proportional to the average node degree in that region. Leveraging this indirect sign of connectivity enables our proposed framework to effectively scale with graph size.

After describing the design of WalkAbout, we demonstrate the capabilities of WalkAbout by applying it to three major OSNs (i.e., Flickr, Twitter, and Google+) and obtaining a coarse view of their connectivity structure. In addition, we examine the “meaning” of the regions obtained by WalkAbout, demonstrate that users in the identified regions exhibit common social attributes, and illustrate how the communities that are obtained by running a popular community detection method on these OSNs stack up against the WalkAbout-discovered regions and their meaning.

Categories and Subject Descriptors
H.2.8 [Database Applications]: Data mining; E.1 [Data Structures]: Graphs and networks

General Terms
Algorithms, Performance

1. INTRODUCTION
Large-scale, networked systems such as the World Wide Web or Online Social Networks (OSNs) can be represented as graphs where nodes represent individual entities, such as web pages or users, and directed or undirected edges represent relations between these entities, such as interaction or friendship between users [12, 16, 21]. Characterizing the connectivity structure of such a graph, in particular at scale, often provides deeper insight into the corresponding networked system and has motivated many researchers to analyze graph representations of large networked systems (e.g., [1]).

It is often very useful to obtain a coarse view of the connectivity structure of a huge graph that shows a few major tightly connected components or regions of the graph along with the inter- and intra-component connectivity. Such a regional view also enables a natural top-down approach to the analysis of large graphs, where one first examines the regional connectivity of a huge graph and then zooms in to individual regions to explore their structure in further detail. However, capturing a regional view of a huge graph is a non-trivial task that existing tools and techniques are not able to achieve. While many techniques exist for graph clustering [23, 5], graph partitioning [11], and community detection [3, 19, 8], these approaches do not work well for discovering coarse regional views in very large graphs. These methods usually scale poorly, force regions to have similar size, or find communities that are too small.

This paper presents a simple top-down framework, called WalkAbout, to identify tightly connected regions in a large unknown graph and subsequently characterize the regional view of its connectivity structure. The main idea is to leverage the behavior of an army of short random walks (RW) on a graph to identify nodes that are located in the same region. When the random walks are longer than the mixing time of an individual region and shorter than the mixing time of the overall graph, the ratio of node degree to expected number of visits is proportional to the edge density of that region. We refer to this quantity as the degree/visit ratio (\(dvr\)). If individual regions in a graph have different edge densities and shorter mixing times than the entire graph, we can leverage the \(dvr\) “signal” to identify the regions, their corresponding nodes and their intra- and inter-region connectivity. The main novelty of WalkAbout is to leverage this indirect sign of connectivity to identify tightly connected nodes in a region. This leads to a very scalable method: in a graph with \(|V|\) nodes, \(|E|\) edges, and a regional mixing time of \(wl\), WalkAbout requires only \(O(wl \times |E|)\) time and \(O(|V|)\) space. A few parameters in WalkAbout enable one to explore different
aspects of the regional connectivity in order to produce the outcome with the desired resolution.

In our empirical evaluation, we apply WalkAbout to three major OSNs: Flickr, Twitter and Google+. Compared to Louvain [3], the gold standard for scalable community detection, WalkAbout runs faster and finds larger, coarser regions. Most communities discovered by Louvain can be mapped to a single one of WalkAbout’s regions, suggesting that WalkAbout is providing a higher-level view of the network than Louvain. Finally, we analyze the regions in Flickr and find that different regions discovered by WalkAbout correspond to different interest groups, providing a meaningful coarse view of this OSN.

The remainder of our paper is organized as follows. Section 3 explores the behavior of short random walks and dvr on graphs with a single region. Section 4 extends this analysis to multiple region graphs and motivates using dvr for region identification. In Section 5, we present the full details of WalkAbout, our step-by-step framework for identifying regions in large graphs. To demonstrate and evaluate WalkAbout, we apply it to three major OSNs in Section 6, compare the results to Louvain in Section 7, and demonstrate that the regions discovered by WalkAbout are meaningful in Section 8. We conclude in Section 9 with a summary of our main contributions and a discussion of future work.

2. BACKGROUND & RELATED WORK

We begin with a brief overview of related work in community detection and graph partitioning.

Most methods work by optimizing an objective function. Since this is typically NP-hard, greedy or heuristic methods are usually necessary. One of the most popular metrics for community detection is modularity, which relates the number of edges within a cluster to the expected number for a random graph. For optimizing modularity, one of the most scalable and effective algorithms is the Louvain method [3]. The Louvain method greedily assigns nodes to communities based on their local connectivity, then coarsens the graph by replacing each community with a single node. This procedure repeats until it reaches a local optimum of modularity. However, in most real-world graphs, modularity tends to favor smaller communities of around 100 nodes [14]. Other measures such as conductance also tend to favor small clusters in real-world graphs, limiting their effectiveness at describing high-level structure.

Community detection methods based on RWs and “flows” have been proposed as well [22, 19, 20]. These use RWs or the associated transition matrix to compute some kind of distance or similarity relationship between each pair of nodes. However, even computing and storing sparse pairwise information is usually too expensive on large graphs with millions of nodes.

Graph partitioning or global clustering techniques [11] adopt a top-down approach, dividing a graph into strongly connected partitions and optionally recursing within each partition to obtain the desired granularity [7, 11]. While this does discover larger regions than the bottom-up approaches, these regions may or may not faithfully represent the overall graph structure. For example, methods that optimize the popular normalized cut criterion tend to produce regions of approximately equal size, even when this leads to poorly separated regions. Furthermore, some approaches require specifying seed instances for each cluster [2] or the total number of partitions, both of which can be difficult to determine a priori. Finally, many of these techniques, including spectral clustering [10], do not scale with graph size and often require a complete snapshot of the target graph or its adjacency matrix.

3. THE BEHAVIOR OF MANY SHORT RWS

Random Walks (RW) are a well-known technique for sampling graphs. A RW on a graph starts from an arbitrary node and at each step moves to a randomly chosen neighbor of the current node. Consider a graph $G = (V, E)$ where $V$ and $E$ denote the set of graph vertices and edges, respectively. In an undirected, connected, and non-bipartite graph, the probability that a sufficiently long RW would be at a particular node $x$ converges to $\frac{1}{\deg(x)}$ [15]. The mixing time $T_{\text{dvr}}(\epsilon)$ of a graph is the walk length at which the probability of being at each node is within $\epsilon$ of the stationary distribution. In this paper, we will use this term somewhat informally, without specifying a particular value of $\epsilon$.

Suppose we run $|V|$ RWs in parallel, one starting at each node. Let $V(x, wl)$ denote the expected number of RWs that are at a particular node $x$ after $wl$ number of steps (e.g., walk length of $wl$). Since one RW is started at each node, $V(x, 0) = 1$. For other values of $wl$, we can define $V(x, k)$ inductively:

$$V(x, 0) = 1$$
$$V(x, wl) = \sum_{n \in \text{Neighbors}(x)} \frac{V(n, wl-1)}{\deg(n)} \quad \text{for } wl > 0$$

This function can be computed iteratively with complexity $O(|E|wl)$. As $wl$ reaches the mixing time, $V(x, wl)$ converges to $|V|^{-\deg(x)}$. Hence, when $wl$ is large, the following holds for all nodes:

$$\frac{\deg(x)}{V(x, wl)} \approx 2 \frac{|E|}{|V|}$$

We refer to the fraction $\frac{\deg(x)}{V(x, wl)}$ as the degree/visit ratio ($\text{dvr}$). From (2), the $\text{dvr}$ converges to the average degree of the graph.

In practice, estimating the mixing time for an arbitrary graph is a known hard problem. In this section, we will explore the dependence of $\text{dvr}$ on $wl$ through simulations on different synthetically-generated graphs. The graphs are generated randomly by selecting the range of node degrees, the distribution of node degrees across this range, and then randomly connecting the nodes until all half-edges are connected. For each simulation, we show a normalized histogram of $\text{dvr}$ values across all nodes, which represents the empirical distribution of $\text{dvr}$ values for that simulation.

Effect of Walk Length: Figure 1(a) shows the evolution of the $\text{dvr}$ histogram as we increase walk length over a generic random graph. As the walk length increases, the variation in $\text{dvr}$ across different nodes decreases, leading to the formation of a narrower peak in the histogram. As $wl$ reaches the mixing time, the probability of visiting each node becomes approximately proportional to its degree.

Effect of Mixing Time: To explore the effect of mixing time on the $\text{dvr}$ histogram, we show in Figure 1(b) the evolution of the $\text{dvr}$ histogram for a small-world graph as we increase the level of clustering (and thus the mixing time) for a particular walk length ($wl = 20$). As the mixing time
becomes longer, the variation in $dvr$ values increases because the RWs are farther from convergence.

Effect of Average Node Degree ($E$): Figure 1(c) presents the effect of average node degree (i.e., changing $|E|$ when $|V|$ is fixed) on the shape of the $dvr$ histogram at a given walk length ($|wl| = 20$). Increasing the average node degree shifts the corresponding peak to higher $dvr$ values. It is worth noting that the placement of each peak is in perfect agreement with the average degree of each graph.

Effect of Minimum Node Degree: Figure 1(d) shows the contribution of low degree nodes to the shape of the $dvr$ histogram by plotting the histogram only for nodes whose degree is larger than a threshold $D_{min}$. We find that higher degree nodes show less variation in $dvr$ than low degree nodes, i.e., filtering low degree nodes leads to a sharper peak in the histogram. Figure 1(e) depicts the evolution of summary $dvr$ values for all nodes across two groups of nodes with different degree which shows that the range of $dvr$ is inversely proportional with node degree and rapidly decreases with the walk length. This property is due to the fact that higher degree nodes are averaging over more neighbors in each update of $V(x, wl)$, thus reducing the variation.

4. DETECTING REGIONS IN A GRAPH

To infer a coarse view of graph connectivity, we assume that each graph consists of a number of weakly inter-connected regions where individual regions have varying edge density. We have no a priori knowledge of either the number of regions or their relative size and make no assumptions about the precise nature of the inter-region connectivity or intra-region connections.

4.1 The Key Idea

Our approach is based on a simple idea – leverage the behavior of RWs that are shorter than the mixing time of the graph to identify nodes in each of the graph’s regions. To this end, consider RWs that start from randomly selected nodes of a graph $G = (V, E)$ that has multiple regions. Based on our discussion in Section 3, the fraction of RWs that starts in region $i$ ($G_i = (V_i, E_i)$) of the graph is equal to the fraction of nodes in that region (i.e., $|V_i|/|V|$). If the length of those RWs is approximately equal to the mixing time of regions $G_i$, a majority of RWs will remain within that starting region, and for all practical purposes, we can view the different regions of the graph as disconnected partitions. Thus, we can use Equation (2) to determine the value of the $dvr$ ratio to which node $x$ in region $i$ converges to as follows:

$$dvr_i(x) = \frac{deg(x)}{E[V(x, wl)]]} = \frac{2 \times |E_i|}{|V_i|}$$

Equation 3 shows that the degree-to-visit ratio for nodes $x$ in region $i$ equals $2\times|E_i|/|V_i|$, which is the average node degree for region $i$. Therefore, if regions of the graph have different average node degrees, the $dvr_i(x)$ values for nodes in each region converge to a different $dvr$ value, i.e., form a peak at a different location in the $dvr$ histogram across all nodes. We can represent each region with its associated non-overlapping range of $dvr$ values and then map visited nodes to a region based on their $dvr$ values. Furthermore, as discussed earlier, other key connectivity features of a region $i$ (e.g., mixing time and size) affect the shape of the corresponding peak.

As the length of the RWs increases beyond the mixing time of individual regions, the RWs are likely to leave their starting regions and contribute to the number of visits for nodes in other regions of the graph. This in turn decreases the gap in the $dvr_i(x)$-values for different regions and the $dvr$ values for all nodes converge to a single value (determined by equation (2)) as soon as the walk length of the RWs agrees approximately with the mixing time of the entire graph. Therefore, the separation between peaks in the $dvr$ histogram that are associated with different regions of a graph is a transient phenomenon that occurs for RWs whose walk lengths are between region-specific mixing times and the mixing time for the entire graph. The more pronounced the regions, the larger the gaps between the mixing times of individual regions and the entire graph, which in turn translates to a longer transient phase and simplifies the detection of different regions. In a nutshell, the similarity in $dvr$ value serves as a promising indirect signal that reveals a tight connectivity among a group of nodes in a graph. The indirect nature of the $dvr$ signal coupled with the ability to efficiently obtain $dvr$ values using short random walks enables our approach to scale with graph size.

4.2 Validation with Synthetic Graphs

Next we use synthetic graphs to demonstrate how our basic idea can reveal (or decode) the regional connectivity features within a graph. To this end, we consider a graph $G$ with two regions, $R_0$ and $R_1$, both with 4K nodes and random connectivity and average degree of 70 and 60, respectively. We connect these two regions with $b$ bridge connections, where each bridge connection is between a pair of random nodes from these regions, and its default value is $b=10k$. In essence, the value of $b$ controls the inter-region connectivity and thus the mixing time of the entire graph.

To illustrate the effect of regional connectivity features on the shape of the $dvr$ histogram, we keep region $R_0$ fixed and systematically change features of $R_1$ and the value of $b$.

Figure 2(a) shows the evolution of the $dvr$ histogram as we vary the average node degree in $R_1$ between 50 and 66. We observe that as the average degree of $R_1$ increases, the corresponding peak gradually moves to higher $dvr$ values and blends into the peak for $R_0$ until individual peaks are no longer distinguishable. Figure 2(b) shows how varying...
the size of $R_i$ from 1K to 8K nodes affects the shape of the 
$dvr$ histogram when all other parameters remain constant. 
Increasing the size of a region proportionally increases the 
number of RWs that starts from that region which in turn 
leads to a proportionally larger peak. Since we normalize 
$dvr$ and plot the PDF, the peak corresponding to $R_0$ de-
creases in size. Finally, Figure 2(c) illustrates the effect of 
increasing the number of bridge edges (or bridge width) be-
tween the two regions from 5K to 50k. We note that as the 
bridge width increases, the two peaks gradually merge and 
become less and less distinguishable. This is due to the fact 
that increasing bridge width decreases the mixing time of the 
entire graph and thus shrinks the transition phase where the 
peaks for two regions can be clearly identified.

In summary, these examples illustrate that the behavior of 
many short RWs on a single graph can be extended to multi-
region graphs as long as the mixing time of the entire graph 
is sufficiently larger than the the mixing time of individual 
regions.

5. WALKABOUT

In this section we present WalkAbout, our proposed method 
for inferring and examining a regional (i.e., coarse) view of 
connectivity for large graphs. We first discuss some of the 
basic challenges in designing such a methodology and then 
describe our approach and how it addresses these challenges.

5.1 Basic Challenges

The behavior of many short RWs on a large graph moti-
velates the idea of using the similarity of $dvr$ values to iden-
ify individual regions of a graph where regions are repre-
sented as a collection of nodes with non-overlapping ranges 
of $dvr$ values. To implement this idea in practice, a num-
ber of challenges arise. First, we recall that the variation 
of $dvr$ values across nodes with degree $d$ in a given region 
decreases monotonically while the median value converges 
towards the average node degree of the region. More impor-
tantly, the degree of variation and its rate of convergence is 
inversely proportional to the node degree $d$, i.e., $dvr$ values 
of higher degree nodes exhibit smaller variations and con-
vergence faster than lower degree nodes. The typically large 
fraction of low degree nodes in big graphs coupled with the 
widely variation and slower convergence rate of their $dvr$ val-
ues make it difficult to accurately associate a set of nodes 
with their corresponding region. This problem is further 
exacerbated by the fact that different regions may have a 
different mixing time and overlapping ranges of $dvr$ values.

5.2 Main Steps of WalkAbout

Given a large graph $G(V,E)$, the goal of WalkAbout is 
to identify the number of regions, map all nodes to their 
corresponding region, and determine the inter- and intra-
region connectivity (i.e., fraction of edges that are connect-
ing nodes in different regions or the same region). We call 
such representation of a large graph a regional (or coarse) 
view of the graph. To overcome the above-mentioned chal-
 lenges, WalkAbout identifies individual regions in two steps. 
First, it identifies a “core” component for each region. Such 
a component consists of a collection of high degree nodes in 
that region based on the similarity of their $dvr$ values. Sec-
ond, it considers each of these core components, views their 
elements as “anchors” and maps the remaining low degree 
nodes to the various regions based on the nodes’ relative 
reachability from each core. This approach can effectively 
cope with the variations of the $dvr$ values for low degree 
nodes and is less sensitive to the walk length. The WalkA-
bout technique comes with a set of parameters/options that 
enable the exploration of the regional connectivity of a graph 
and support experimentation with different coarse views of 
a graph. In the following, we describe the five main steps of 
the WalkAbout technique.

4) Determining $dvr$ values for individual nodes: Given the $dvr$ values 
of different nodes, our goal is to group nodes with similar 
$dvr$ values and use them as the core elements for the cor-
responding region. To this end, we bin the nodes based on 
their $dvr$ values and generate a histogram to identify the 
most common values (i.e., “peaks”) which in turn suggest 
the existence of different regions. To reduce the noise that 
the wide variation of $dvr$ values for low degree nodes intro-
duces, we first filter out all nodes whose degree is smaller 
than a threshold $D_{min}$. In fact $D_{min}$ is a parameter that 
can be used to control the visibility of nodes that are under 
possible consideration for being selected as core elements. 
It provides a knob for examining the trade-offs that result 
from increasing the level of noise caused by a larger number 
of low degree nodes (i.e., small $D_{min}$ values) – allowing for 
more noise typically results in the identification of more but 
less reliable core elements and hence regions. Next, while the 
$dvr$ value for higher degree nodes are significantly more 
reliable, these nodes may not have a profound impact on 
the shape of the histogram due to the often small fraction of 
high degree nodes. We deal with this issue by introducing a 
bias towards the $dvr$ values of high degree nodes. In partic-
ular, for each high degree node, we multiply its $dvr$ value by 
its node degree. In effect, we simply increase the frequency 
of the $dvr$ values of the high degree nodes proportional to 
their node degree and as a result, the modified histogram is
in general more amenable to reveal the presence of reliable regions via the existence of more pronounced peaks.

3) Identifying core of a region from the histogram: Identifying regions from a *dvr* histogram requires (i) determining a proper walk length that generates the best histogram, and (ii) detecting the regions from the resulting histogram. To deal with item (i), we progressively increase the walk length and repeat step (1) and (2) to generate the resulting histogram. We carefully examine the evolution of the histogram as a function of walk length and select the histogram where the peaks are most pronounced and most separated. By definition, such a histogram should be formed when the walk length is close to the mixing time of individual regions. In such a histogram, each peak (i.e., a local maximum that is surrounded by two local minimum values) represents a region’s core whose range of *dvr* values is specified by the *dvr* values of the two minimum values. This heuristic can be viewed as a naive one-dimensional clustering technique. We examine the connectivity among nodes that are part of each core to ensure that they form a connected component. This check also reveals whether the cores of two separate regions with overlapping *dvr* ranges appear as a single peak which makes it difficult to distinguish them from the histogram in the first place. At the end of this step, we have the number of regions and the list of high degree nodes that form the core of each region.

4) Mapping Low-Degree Nodes to Cores: We use the relative reachability of low degree nodes to identified cores in order to map them. To this end, we start *N* RWs from each node where each RW walk continues until it hits a node in one of the cores. Each walk provides a sample of reachability for this node. The node is mapped to the core with the highest reachability. The fraction of RWs that hit the most reachable core indicates our confidence in mapping a node to that region.

5) Producing the regional view: Once nodes in each region of the graph are identified, we determine the edges that are within each region or connecting two different regions. Then we produce a diagram that incorporates all the information about regional connectivity of a graph including (i) a circle represents a region with the area proportional to the log size of the region, (ii) arrows between two region indicate the inter-region connectivity and their width is proportional with the fraction of directed half-edges between two regions. Intra region half-edges are represented with the modularity of a region and thus are not shown in the regional view to keep this less crowded.

5.3 Inferring vs. Exploring Regions

The design of WalkAbout provides several parameters or knobs that can be tuned to explore different coarse views of a given graph. These parameters include the walk length, the *D*\(_\text{min}\)* threshold, and the precise nature of determining how low degrees nodes get mapped to regions (core anchors). In essence, examining the effect of these parameters on the resulting regional views facilitates studying the quality of a given regional view in terms of its robustness to the choices WalkAbout offers to its users. In this sense, WalkAbout can be viewed as a framework for exploring regional connectivity in an interactive manner rather than a technique for producing a single regional view.

It is also important to emphasize that since WalkAbout is not trying to optimize an explicit objective function (e.g., modularity [18], the regional view that results from running WalkAbout for a given graph is not unique. Instead, by harvesting a transient phenomenon, we face a new challenge in the form of deciding on a proper walk length. Our approach to deal with this challenge is to gain an understanding of the sensitivity of a resulting regional view to the choice of the walk length to minimize potential mistakes at each step.

By varying the *D*\(_\text{min}\) parameter, we are enable to explore the trade-off between level of coarsening and the accuracy of the regional view. Large values of this parameter typically result in few but reliable regions (i.e., coarse and stable view), while smaller values of *D*\(_\text{min}\) produce in general many more but less reliable regions (i.e., fine but unstable views). Alternatively, *D*\(_\text{min}\) can be set based on domain knowledge to only include nodes that are considered central for a given context. For example, in an OSN graph, nodes with degree larger than 500 or even 1000 may be viewed as core nodes. In this paper, we focus primarily focus on the application of WalkAbout to OSN and set *D*\(_\text{min}\) to 500.\(^2\)

We have developed WalkAbout as an interactive tool with GUI that allows users to slice the histogram and observed the resulting regional view. We plan to make the tool publicly available in the next few months.

6. WALKABOUT IN ACTION

In this section, we use our proposed technique to characterize coarse views of large popular OSNs such as Flickr, Twitter, and Google+. In the process, we not only demonstrates the key features and capabilities of our technique, but also show what sort of coarse views WalkAbout produces for the well-known OSNs.

6.1 Datasets and methodology

In the following, we rely on anonymized snapshots of the largest connected component (LCC) of the connectivity structure for Flickr (FL) that were captured by Mislove et al. [16], a snapshot of the Twitter (TW) social graph that was collected by Kwak et al. [13], and a snapshot of Google+ (G+) from a recent study by Gonzalez et al. [9]. Table 1 summarizes the main characteristics of these snapshots.

Table 1: Characteristics of LCC snapshots of target OSNs

<table>
<thead>
<tr>
<th></th>
<th>FL</th>
<th>TW</th>
<th>G+</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nodes</td>
<td>1.6M</td>
<td>41.6M</td>
<td>51.7M</td>
</tr>
<tr>
<td>Edges</td>
<td>31.1M</td>
<td>1.4E8M</td>
<td>8.6E8M</td>
</tr>
<tr>
<td>Crawl Time</td>
<td>Jan’07</td>
<td>Jul’09</td>
<td>Apr’12</td>
</tr>
</tbody>
</table>

We have examined the effect of *D*\(_\text{min}\) on the *dvr* histogram and our findings are reported in the related technical report.

\(^2\)We have examined the effect of *D*\(_\text{min}\) on the *dvr* histogram and our findings are reported in the related technical report.
in each core) and the corresponding regions (the percentage of total nodes and edges, average degree and modularity), and (in) a sketch the regional view of the OSN.

We refer to the collection of specified values for the Walk-About parameters, namely $D_{min}$ and $wl$, as the target setting. In particular, we used $D_{min} = 500$ throughout this analysis. To examine the robustness of our results to different choices of $D_{min}$ values, we repeated our analysis with $D_{min}$ values that are 10% larger or smaller and observed no significant differences. For a more detailed account of this robustness analysis, refer to our related technical report [17].

### 6.2 OSNs and their regional views

**Regional View of Flicker (FL):** Figure 3(a) shows the evolution of $dvr$ histogram for a FL snapshot as a function of walk length around the selected target setting ($wl = 30$, $D_{min} = 500$). We observe that $wl = 30$ reveals the largest number of pronounced peaks; i.e., a total of five peaks. Figure 3(b) shows the shape of $dvr$ histogram at our selected target setting for FL ($wl = 30$, $D_{min} = 500$) where the five major peaks are marked and their associated ranges of $dvr$-values are colored. Note that regions $R_3$ and $R_4$ could have been considered as a single region. However, because of the observed dip around $dvr = 35$, we split that peak into two regions. We later discuss the effect of this decision. Due to their small sizes and to keep the number of regions within limits, we did not consider the many small peaks in the middle of the histogram.

Table 2 summarizes the key features of the five identified cores and their corresponding regions. We observe that the cores include between 500-4000 nodes and collectively contain less than 1% of nodes of the graph. Except for $R_1$, they are all of similar size. The resulting regions are very imbalanced, with $R_0$ containing more than 92% of all nodes and having average degree of 11.9 and modularity of 0.4. The other regions are very small and contain only some 0.2%-4% of all nodes. However, regions $R_2$ and $R_3$ have a high average degree and thus include a much larger fraction of edges. At the same time, regions $R_2$ and $R_4$ have a much higher modularity than $R_0$. All the identified cores and regions form connected components. Figure 3(c) sketches the regional view of the FL structure. This figure shows that for all practical purposes, regions $R_3$ and $R_4$ are weakly connected to the other three regions. We recall that these two regions are created as a result of splitting the right most peak of the $dvr$ histogram into two parts. Given their strong inter-connectivity, an option would be to merge these two regions together and consider them as a single region, thus producing a yet coarser view of the FL connectivity structure.

**Regional View of Twitter (TW):** Figure 5(a) depicts the evolution of the $dvr$ histogram for the TW structure as a function of $wl$ where $D_{min} = 500$. We observe that the transition phase for the formation of peaks for different regions is rather short, between $wl$ values of 14 and 22. We select $wl = 18$ as our target setting as it reveals the most clear set of peaks in the histogram. Figure 4(b) depicts six peaks in the $dvr$ histogram at our target setting.

Table 3 summarizes the main characteristics of the identified cores and their corresponding regions. We observe the cores have between 1.9K and 275K nodes. There are two large ($R_1$ and $R_2$), two small ($R_0$ and $R_5$), and two tiny ($R_4$ and $R_5$) regions. The regions generally exhibit low modularity ($\leq 0.4$). The low level of modularity for regions in TW indicates that regions do not exhibit tight internal connectivity. An interesting fact about the two tiny regions is that they have an order of magnitude larger average degree than the other regions but still exhibit the same modularity. Figure 4(c) depicts the resulting regional view for the TW structure and reveals that regions $R_1$ and $R_2$ have strong mutual connectivity and play a central role in the graph. $R_0$ is connected to $R_1$ and $R_2$ from one side while $R_4$ and $R_5$ form a chain structure and connect to the rest of the regions primarily through $R_2$.

**Regional View of Google+ (G+):** Figure 5(a) depicts the evolution of the $dvr$ histogram for the G+ graph as we change $wl$. The histogram which most clearly reveals different regions is formed around $wl = 20$. Therefore, we select this $wl$ as our target setting. The corresponding histogram is shown in Figure 5(b) and reveals the existence of six distinguishable peaks. While the regions $R_4$ and $R_5$ result from rather small peaks, we still use them as cores because they are clearly separated from other peaks and also have a large average degree.

Table 4 summarizes the main features of the identified cores and regions. We observe that the core sizes vary between 74 and 218k which is much more skewed compared to the other OSNs. These cores lead to a dominant region $R_0$, a moderate-sized region $R_1$, and four tiny regions. All regions except for $R_0$ exhibit a rather high modularity (0.6-0.8). Figure 5(c) plots the regional view of the connectivity structure for G+. We observe that $R_4$ and $R_5$ are tightly inter-connected but have a weak connectivity to the other regions. The other four regions have a moderate chain-like inter-connectivity structure of the form $R_2$-$R_3$-$R_1$-$R_0$.

### 6.3 Lessons learned

The obtained regional views of the connectivity structures of some of the most popular OSNs provide a novel and useful abstraction of the large-scale real-world systems. They provide a manageable high-order view of how nodes are mapped into various regions of different sizes, along with a quantita-
A common observation from applying WalkAbout to the three OSNs is that separate regions (peaks) with close-by dvr values tend to have stronger inter-region connectivity than regions that result from clearly separated peaks in the vdr histogram. Such behavior is to be expected for real-world graphs. For one, our approach for mapping high degree nodes to cores based on slicing peaks in the histogram is ambiguous for high degree nodes whose dvr values are close to the border value of a region. Moreover, the size of a region and its mixing time can vary widely across different regions of a large graph. This in turn makes the selection of a proper walk length challenging.

For example, a particular walk length that is close to a region’s mixing time and thus clearly reveals the associated peak in the vdr histogram could be too long for other regions. This behavior can cause some of the RWs of other regions to leave their starting points and move to other close-by regions. The fraction of such “misbehaving” RWs depends on the walk length and the relative connectivity between starting and neighboring regions. Both of the above factors tend to decrease the gap between the dvr ranges of close-by regions proportional to their pairwise connectivity. However, given the coarse resolution of the considered regional views of a graph, the resulting ambiguities do not significantly impact the value that can be derived from examining such coarse views of large-scale graphs.

Also note that the number of peaks that appear in a vdr histogram changes with the walk length which, in turn, can change the perspective of what peak size should be considered to be significant. Our focus here has been on considering only a handful of regions so that the resulting regional views are manageable. WalkAbout is clearly an interactive framework and can be used to identify a different number of regions and examine how such selections affect the characteristics of the resulting regional views.

7. REGIONS & COMMUNITIES

Community detection in graphs is a commonly used technique that can also be viewed as providing a coarse view of a graph (i.e., community-level instead of regional-level view). Community detection techniques typically group nodes into tightly connected groups, called a community, based on an objective function (e.g., modularity) and present characteristics of the detected communities without emphasis on the inter-community connectivity. In this section, we compare and contrast the regional view that WalkAbout produces with the community view of a large graph. Given the similarity between the notion of a “community” and a “region”, and the popularity of applying community detection techniques for graph analysis, this comparison helps us relate the regional view of the graph with a related concept (i.e., community) that is widely used. To this end, we have to run a community detection technique on our large target graphs. Unfortunately, most of the commonly-used community detection techniques do not scale to graphs with more than tens of millions of nodes [6], or require the number of communities as an input (e.g., Metis [11]), or recursively partition the graph into balanced communities that may not lead to the most tightly connected communities [11]. Due to these limitations, we use the Louvain community detection technique [3] that implements a greedy method to optimize the “modularity” of identified partitions. The Louvain technique is often considered to be the gold standard for scalable community detection and has a publicly available and robust implementation.

We applied Louvain to our targeted OSN structures and identified 28K, 39K, and 24K communities of various sizes in FL, G+, and TW, respectively. Importantly, these results show that the number of communities in these graphs are several orders of magnitude larger than the number of regions. This large number of communities implies that the
Interestingly, we find that in all three OSNs, the top 1000 nodes with the highest degree of the communities clearly have an overlap with the obtained regions. The only exceptions to this observation are regions distinguishable based on the characteristics of their communities. For high-level analysis (e.g., Louvain), the high-degree nodes exhibit characteristics very similar to the WalkAbout-derived regions. These results suggest that the large communities that are needed for accommodating high-degree nodes exhibit characteristics very similar to the WalkAbout-identified regions.

Mapping communities to regions: To further explore the relationship between the community- and regional-level views of these graphs, we map individual Louvain communities to the identified regions for the same graph. In particular, for each community $c$, we determine the region where each node of this community is located and identify the region $R$ that contains a majority of nodes in that community. Then community $c$ is mapped to that region $R$ that hosts a majority of its nodes, and the confidence for this mapping is equal to the fraction of $c$’s nodes that are located in $R$. Tables 6 through 7 summarize the number of communities that are mapped to the individual regions of each OSN. In the extreme case, if the nodes in each community are randomly located in different regions, then all communities are mapped to the largest region(s) with a confidence equal to the region’s relative size. We observe that the mapping confidence for 75% of the communities in every single region is 100%, and for 90% of communities, all but one small region in FL ($R_4$) has a mapping confidence higher than 80%. Even for the large communities with more than 1K nodes, the mapping confidence for 90% of them is larger than 80% for all regions of all OSNs except for TW, where it is 60%. These results clearly demonstrate that the vast majority of nodes in most communities is mapped into a single region. This in turn suggests that a region can be viewed as a collection of connected communities.

Per-region analysis of communities: Figure 7 uses boxplots to summarize the distribution of modularity, size, and average degree across all communities that are mapped to each region of individual OSNs. These figures illustrate that there does not appear to be a strong correlation between the modularity of communities in a region and the modularity of the entire region. This observation is explained by the fact that the modularity of a region depends, among other factors, on the inter-community connectivity. We also observe that in general, there is no significant difference in the modularity, size and average degree of the communities that are mapped to each region, i.e., regions are not generally distinguishable based on the characteristics of their communities. The only exceptions to this observation are regions communities attracting significantly more nodes than others. Moreover, both the size (15K-359K for FL, 72K-22M for TW, and 336K-16M for G+) and the modularity of these few communities (0.48-0.75 for FL, 0.28-0.78 for TW, and 0.35-0.89 for G+) are comparable with typical values for the WalkAbout-derived regions. These results suggest that the large communities that are needed for accommodating high-degree nodes exhibit characteristics very similar to the WalkAbout-identified regions.

Graph connectivity at the community level is still too complex for high-level analysis (e.g., determining the full picture of inter-community connectivity.

Figure 6 presents the summary distribution of the main features (modularity, size and average degree) across all regions and all communities associated with each each OSN. To examine the effect of community size, we have also included the results where we only consider the large communities that consist of 1000 or more nodes. We observe that communities are typically more than four orders of magnitude smaller than regions. However, size-wise, the largest communities clearly have an overlap with the obtained regions. While the modularity of communities is typically higher than the modularity of regions, this gap is more pronounced in less clustered graphs (e.g., TW) than in more clustered graphs like FL and G+.

To gain more insight into connectivity-related features, we examine the placement of the 1000 nodes with the highest degree in each region across the different communities. Interestingly, we find that in all three OSNs, the top 1000 nodes are located in 5 or 6 communities, with some of those communities attracting significantly more nodes than others. Moreover, both the size (15K-359K for FL, 72K-22M for TW, and 336K-16M for G+) and the modularity of these few communities (0.48-0.75 for FL, 0.28-0.78 for TW, and 0.35-0.89 for G+) are comparable with typical values for the WalkAbout-derived regions. These results suggest that the large communities that are needed for accommodating high-degree nodes exhibit characteristics very similar to the WalkAbout-identified regions.

Mapping communities to regions: To further explore the relationship between the community- and regional-level views of these graphs, we map individual Louvain communities to the identified regions for the same graph. In particular, for each community $c$, we determine the region where each node of this community is located and identify the region $R$ that contains a majority of nodes in that community. Then community $c$ is mapped to that region $R$ that hosts a majority of its nodes, and the confidence for this mapping is equal to the fraction of $c$’s nodes that are located in $R$. Tables 6 through 7 summarize the number of communities that are mapped to the individual regions of each OSN. In the extreme case, if the nodes in each community are randomly located in different regions, then all communities are mapped to the largest region(s) with a confidence equal to the region’s relative size. We observe that the mapping confidence for 75% of the communities in every single region is 100%, and for 90% of communities, all but one small region in FL ($R_4$) has a mapping confidence higher than 80%. Even for the large communities with more than 1K nodes, the mapping confidence for 90% of them is larger than 80% for all regions of all OSNs except for TW, where it is 60%. These results clearly demonstrate that the vast majority of nodes in most communities is mapped into a single region. This in turn suggests that a region can be viewed as a collection of connected communities.

Per-region analysis of communities: Figure 7 uses boxplots to summarize the distribution of modularity, size, and average degree across all communities that are mapped to each region of individual OSNs. These figures illustrate that there does not appear to be a strong correlation between the modularity of communities in a region and the modularity of the entire region. This observation is explained by the fact that the modularity of a region depends, among other factors, on the inter-community connectivity. We also observe that in general, there is no significant difference in the modularity, size and average degree of the communities that are mapped to each region, i.e., regions are not generally distinguishable based on the characteristics of their communities. The only exceptions to this observation are regions communities attracting significantly more nodes than others. Moreover, both the size (15K-359K for FL, 72K-22M for TW, and 336K-16M for G+) and the modularity of these few communities (0.48-0.75 for FL, 0.28-0.78 for TW, and 0.35-0.89 for G+) are comparable with typical values for the WalkAbout-derived regions. These results suggest that the large communities that are needed for accommodating high-degree nodes exhibit characteristics very similar to the WalkAbout-identified regions.

Mapping communities to regions: To further explore the relationship between the community- and regional-level views of these graphs, we map individual Louvain communities to the identified regions for the same graph. In particular, for each community $c$, we determine the region where each node of this community is located and identify the region $R$ that contains a majority of nodes in that community. Then community $c$ is mapped to that region $R$ that hosts a majority of its nodes, and the confidence for this mapping is equal to the fraction of $c$’s nodes that are located in $R$. Tables 6 through 7 summarize the number of communities that are mapped to the individual regions of each OSN. In the extreme case, if the nodes in each community are randomly located in different regions, then all communities are mapped to the largest region(s) with a confidence equal to the region’s relative size. We observe that the mapping confidence for 75% of the communities in every single region is 100%, and for 90% of communities, all but one small region in FL ($R_4$) has a mapping confidence higher than 80%. Even for the large communities with more than 1K nodes, the mapping confidence for 90% of them is larger than 80% for all regions of all OSNs except for TW, where it is 60%. These results clearly demonstrate that the vast majority of nodes in most communities is mapped into a single region. This in turn suggests that a region can be viewed as a collection of connected communities.
Figure 7: Modularity of Louvain communities mapped to different WalkAbout regions

Figure 8: The comparison of the execution time for different techniques.

$R_3$, $R_4$ and $R_5$ in G+ that contain communities with a significantly higher and more homogeneous modularity, larger size and higher average degree. This is intriguing since larger size or higher node degree could lead to lower modularity in a single community. These findings suggest that identifying individual regions by merging communities in a bottom-up fashion (using modularity) is in general challenging. Alternatively, a top-down approach to region detection such as WalkAbout shows more promise.

Comparing Run-time: Finally, we compare the run times of WalkAbout and the Louvain community detection technique on an Intel X5650 2.66GHz computer with 72GB RAM which is sufficient to hold the entire graph in memory. Figure 7 shows the comparison of the run time per individual technique over each OSN using log scale for the x-axis. We further split the run time of WalkAbout into two components: (i) the calculation of the $dvr$ values for high degree nodes to detect cores and (ii) mapping of low-degree nodes to those cores. These results show that the run times of both techniques are similar over small graphs (e.g., 10 seconds for FL). However, as the graph size increases, Louvain requires a significantly longer run time and the gap between WalkAbout and Louvain seems to be widening. We also recall that for graphs of the size of these OSNs, many popular community detection or clustering techniques (including spectral clustering [4]) quickly run into scalability issues and cannot be used at all.

8. A NEW KIND OF VALIDATION

So far we have primarily focused on the connectivity features of regions and how they are aligned with smaller entities in a large graph such as communities. Since regions are not derived based on an objective function, there is no obvious way to validate/examine their accuracy. To tackle the challenging problem of “validation” of WalkAbout -derived regions, we conduct a case to investigate whether users in each identified region exhibit similar social attributes that act as the underlying factors for the formation of the region.

8.1 Are regions meaningful?

Our ability to answer the posed question depends on the availability of semantically-rich metadata that contains social context. However, given adequate metadata, answering the above question will shed light on whether an identified region represents a meaningful portion of an OSN. In our case study, we focus on FL because of the availability of rich metadata with social context for this OSN.

More precisely, for our FL snapshot, we have a list of 99K social groups (with their names) where each group consists of a collection of users with common interest. A user can be a member of multiple groups. The name of most groups provides a great deal of information about the groups’ interests or context (e.g., big_and_hot, bigboy, fatboys). Similarly to the mapping of communities to regions (see Section 7), we map each group to a region where most of their users are located. Figure 9(a) and 9(b) shows the summary distribution of mapping confidence for all groups and for the 44K groups with more than 10 users to the five regions in FL, respectively. We observe that groups that are mapped to regions $R_1$-$R_4$ exhibit a very high confidence despite the small size of these regions. The mapping confidence drops for larger groups but it is still a couple of orders of magnitude larger than the relative size of the group. More specifically, regions $R_1$, $R_2$, $R_3$ and $R_4$ make up 1.2%, 4%, 1.8% and 0.2% of nodes in the graph but the typical confidence for their mapped groups is 0.8, 0.9, 0.92, and 0.58, respectively. These results suggest that the social context of each group is likely a driving force for its mapping to these four regions. In contrast, the typical confidence for mapped groups to region $R_5$ is comparable to its relative size. This indicates that social forces discernible from our data may not be primarily responsible for the mapping of groups to region $R_5$. To learn the context of individual regions, we manually examined the names of groups that are mapped to that region. Our examination reveals a very pronounced patterns among group names associated with the following regions: Group names in $R_1$ are mostly related to male nudity and adult content, group
names in $R_2$ are hinting at female nudity and adult content, and group names in both $R_3$ and $R_4$ have a common ethnic attribute, i.e., either have Arabic name or post in Arabic. As expected, group names in $R_0$ do not show a coherent theme.

8.2 Are communities meaningful?

We use the same methodology to examine the “validity” of communities; i.e., checking whether the names of mapped groups to individual communities indicate any common social theme. In the case of regions without any pronounced social theme (e.g., $R_0$), one of their large communities may indeed have a social context whereas for regions with an existing social context (e.g., $R_1$), a community may offer an even more specific context. The large number and diverse size of communities in each graph make it difficult to examine all communities. Since small communities do not provide sufficient information to identify their social theme, we only focus on the three largest communities that are mapped to each region of FL. Careful examination of group names for regions often exhibits a theme that is very similar to the identified theme for the whole region. The only exception is a community in $R_2$ that contains groups with clearly more specific group names. In summary, our preliminary investigations suggest that some large communities that are embedded within a region are not “meaningful” in the sense that they exhibit very diverse social themes that makes them the opposite of a “community.”

9. CONCLUSION & OUTLOOK

In this paper, we present a new scalable framework called WalkAbout for examining and inferring regional views of connectivity for very large graphs and demonstrate its application to three well-known OSNs. Moreover, we conduct a comparison between regional and community-level views of large OSNs and present a case study where we “validate” the individual regions and communities; i.e., examine in detail the available meta-data for social themes that are associated with the obtained groupings of nodes in an OSN and are prime candidates for the root cause(s) behind the formation of these groupings.

The presented design of WalkAbout and the experience we gained from applying it to real-world OSNs suggest a number of extensions and improvements. For one, we plan to explore the recursive application of WalkAbout to identify potential sub-regions within each identified region. In the same vein, we intend to examine how the regional and community-level views of a large graph can inform each other to yield a hybrid approach for a “multi-scale” exploration of the graph’s connectivity (e.g., examining the connectivity between large communities within a given region to obtain a higher-resolution view of graph connectivity). Extending WalkAbout to allow for overlapping regions and collecting semantically rich meta-data that enables the illustrated validations of groupings such as regions, clusters, or communities are other items on our research agenda in this area.

10. REFERENCES