HIT'nDRIVE: Multi-Driver Gene Prioritization based on Hitting Time

Raunak Shrestha[1,2,*], Ermin Hodzic[3,*], Jake Yeung[2,4,*], Kendric Wang[2], Thomas Sauerwald[5], Phuong Dao[6], Shawn Anderson[2], Himisha Beltran[7], Mark A. Rubin[7], Colin C. Collins[2,8], Gholamreza Haffari[9], and S. Cenk Sahinalp[3,10,†]

[1] CIHR Bioinformatics Training Program, University of British Columbia, Vancouver, BC, Canada.
[2] Laboratory for Advanced Genome Analysis, Vancouver Prostate Centre, Vancouver, BC, Canada.
[3] School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.
[4] Genome Science and Technology Program, University of British Columbia, Vancouver, Canada.
[5] Computer Laboratory, University of Cambridge, Cambridge, United Kingdom.
[6] National Center for Biotechnology Information, NLM, NIH, Bethesda, MD, USA.
[7] Weill Cornell Cancer Center, New York, NY, USA.
[8] Department of Urologic Sciences, University of British Columbia, Vancouver, BC, Canada.
[9] Faculty of Information Technology, Monash University, Melbourne, Australia.
[10] School of Informatics and Computing, Indiana University, Bloomington, IN, USA.

[†] E-mail: S. Cenk Sahinalp (cenk@sfu.ca)
[*] These authors contributed equally to this work.

Abstract. A key challenge in cancer genomics is the identification and prioritization of genomic aberrations that potentially act as drivers of cancer. In this paper we introduce HIT'nDRIVE, a combinatorial method to identify aberrant genes that can collectively influence possibly distant "outlier" genes based on what we call the "random-walk facility location" (RWFL) problem on an interaction network. RWFL differs from the standard facility location problem by its use of "multi-hitting time", the expected minimum number of hops in a random walk originating from any aberrant gene to reach an outlier. HIT'nDRIVE thus aims to find the smallest set of aberrant genes from which one can reach outliers within a desired multi-hitting time. For that it estimates multi-hitting time based on the independent hitting times from the drivers to any given outlier and reduces the RWFL to a weighted multi-set cover problem, which it solves as an integer linear program (ILP). We apply HIT'nDRIVE to identify aberrant genes that potentially act as drivers in a cancer data set and make phenotype predictions using only the potential drivers - more accurately than alternative approaches.

1 Introduction

Over the past decade, high-throughput sequencing efforts have revealed the importance of genomic aberrations in the progression of cancer [1]. During the time course of cancer evolution, tumor cells accumulate numerous genomic aberrations, however only a few "driver aberrations" are expected to confer crucial growth advantage - and have potential to be used as therapeutic targets. The identification of these driver aberrations and the specific genes they alter poses a significant challenge as they are greatly outnumbered by functionally inconsequential "passenger"

aberrations which contribute further towards cancer heterogeneity [1, 2].

While several methods for finding drivers of cancer have been described previously, most of them rely on the recurrence frequency of single nucleotide variants with respect to the background mutation rate in a population of tumors [3, 4]. These approaches are restricted to identifying only highly recurrent mutations as driver events. However, recent whole-genome studies have revealed that important genes may be recurrently mutated in only a small fraction of the tumor cohort under study, and can be subtype-specific [5–7]. Furthermore, personalized rare drivers are likely to arise during later stages of tumor evolution and be isolated to a small fraction of tumor cells [8, 9].

Perhaps the first computational method to consider large scale genomic variants as driver events is by Akavia *et al.* [10], which correlates genes with highly recurrent copy number alterations with variation in gene expression profiles within a Bayesian network. Similarly, Masica and Karchin [11] correlate gene mutation information with expression profile changes in other genes, again with no prior knowledge of pathways or protein interactions. Another approach, (Multi) Dendrix [12] aims to simultaneously identify multiple driver pathways, assuming mutual exclusivity of mutated genes among patients, using either a Markov chain Monte Carlo algorithm or integer linear programming (ILP). Finally, MEMo by Ciriello *et al.* [13], identifies sets of proximally-located genes from interaction networks, which are also recurrently altered and exhibit patterns of mutual exclusivity across the patient population. To the best of our knowledge, the first method to link copy number alterations to expression profile changes within an interaction network is by Kim *et al.* [14] which connects specific "causal" aberrant genes with potential targets in a protein interaction network. Similarly, method, PARADIGM [15], computes gene-specific inferences using factor graphs to integrate various genomic data to infer pathways altered in a patient. A more recent tool, HotNet by Vandin *et al.* [16], was the first to use a network diffusion approach to compute a pairwise influence measure between the genes in the (gene interaction) network and identify subnetworks enriched for mutations. TieDIE [17] also uses the diffusion model to identify a collection of pathways and subnetworks that associate a fixed set of driver genes to expression profile changes in other genes. Briefly, the network diffusion approach aims to measure the influence of one node over another by calculating the stationary proportion of a "flow" originating from the starting node, that ends up in the destination node. Since it is based on the stationary distribution, the inferences that can be made by the diffusion model are time independent. In that sense, the diffusion approach is very similar to Rooted PageRank, the stationary probability of a random walk originating at a source node, being at a given

destination node. A final method, DriverNet by Bashashati *et al* [18], also aims to correlate single nucleotide alterations with target genes expression profile changes, but only among direct interaction partners. The novel feature of DriverNet is that it aims to find the "minimum" number of potential drivers that can "cover" targets.

Our Contributions. In this paper we present a novel integrative method that considers potential driver events at the genomic level, i.e. single nucleotide mutations, structural or copy number changes. Our contributions are as follows:

1. We present HIT'nDRIVE, an algorithm that aims to identify "the most parsimonious" set of patient specific driver genes which have sufficient "influence" over a large proportion of outlier genes. HIT'nDRIVE formulates this as a "random-walk facility location" problem (RWFL), a combinatorial optimization problem, which, to the best of our knowledge, has not been explored earlier. RWFL differs from the standard facility location problem by its use of "multi-source hitting time" (or multi-hitting time) as an alternative distance measure between a set of aberrant genes (potential drivers) and an outlier gene. Multi-hitting time generalizes the notion of hitting time [19]: we define it as the expected minimum number of hops in which a random walk originating from any aberrant gene reaches the outlier for the first time (in the human gene or protein interaction network). RWFL problem thus asks to find the smallest (the most parsimonious) set of aberrant genes from which one can reach (at least a given fraction of) all outliers within a user defined multi-hitting time. We believe that applications of RWFL problem may extend beyond its application to driver gene identification - to influence analysis in social networks, disease networks, etc.
2. Since RWFL problem is NP-hard, we estimate the multi-hitting time based on the independent hitting times of the drivers to an outlier, which provides an upper bound on the multi-hitting time. Our experiments show that this estimate works well for the human protein interaction network.
3. More importantly, our estimate enables us to reduce the RWFL problem to a weighted multi-set cover problem, for which we give an ILP formulation. For the specific problem instances we consider, our ILP formulation is solvable exactly by CPLEX in less than two days on a standard PC.
4. Note that hitting time as a measure for influence of one potential driver on an outlier gene is quite different from the diffusion-based measures or the Rooted PageRank: hitting time essentially measures the expected distance/time between a source node and a destination node in a random walk. We argue that hitting time is a better measure to capture the influence of one (driver) node over another as it is (i) parameter free (diffusion model introduces at least one additional parameter - the proportion of incoming flow "consumed" at a node in

each time step), (ii) it is time dependent (while the diffusion model and PageRank measures the stationary behavior) and (iii) it is more robust (w.r.t. small perturbations in the network; see [20]).

5. We also show that, by a simple Monte Carlo method, the hitting time in networks with n nodes that have constant average degree and small diameter (as per the human protein interaction network) can be estimated in $\tilde{O}(n^2)$ time. For computing the hitting time in general networks, alternative methods [21] require to perform a complete matrix inversion, which takes $O(n^{2+c})$ time for some $c > 0.37$.
6. We have applied HIT'nDRIVE to identify genes subject to somatic mutation and copy number changes that potentially act as drivers in glioblastoma cancer. We then used the identified potential drivers to perform phenotype prediction on the cancer data set, solely based on gene expression profiles of small subnetworks "seeded" by the drivers. For that we extended the OptDis method [22] by focusing only on driver-seeded subnetworks and achieved a higher accuracy than the alternative approaches.

2 HIT'nDRIVE Framework

HIT'nDRIVE naturally integrates genome and transcriptome data from a number of tumor samples for identifying and prioritizing aberrated genes as potential drivers. It "links" aberrations at the genomic level to gene expression profile alterations through a gene or protein interaction network. For that, it aims to find the *smallest* set of aberrated genes that can "explain" most of the observed gene expression alterations in the cohort. In other words, HIT'nDRIVE identifies the minimum number of potential drivers which can "cause" a user-defined proportion of the downstream expression effects observed.

HIT'nDRIVE uses a particular "influence" value of a potential driver gene on other (possibly distant) genes based on the (gene or protein) interaction network in use. In order to capture the uncertainty of interaction of genes with their neighbours, it considers a random walk process which propagates the effect of sequence alteration in one gene to the remainder of the genes through the network. As a result, the influence is defined to be the inverse of hitting-time, the expected length (number of hops) of a random walk which starts at a given potential driver gene, and "hits" a given target gene the first time in a (protein or gene) interaction network. More specifically, for any two nodes $u, v \in V$ of an undirected, connected graph $G = (V, E)$, let the random variable $\tau_{u,v}$ denote the number of hops in a random walk starting from u to visit v for the first time. The hitting-time $H_{u,v}$, thus is defined as $H_{u,v} = E[\tau_{u,v}]$ [23].

In order to capture synthetic lethality like scenarios, HIT'nDRIVE also considers multiple aberrated genes as potential drivers. For that, we define the influence value (of a set of potential driver genes on a target) as the inverse of multi(source)-hitting time, i.e., the expectation of the smallest number of hops in one of the random walk processes, simultaneously starting at each one of the potential drivers and ending at a given outlier for the first time. More specifically, let $U \subseteq V$ be a subset of nodes of G and $v \in (V - \{U\})$ be a single node. We thus define the multi(source)-hitting time $H_{U,v}$ as $H_{U,v} = E[\min_{u \in U} \tau_{u,v}]$.

HIT'nDRIVE formulates the process of potential driver gene discovery in terms of the "random-walk facility location" (RWFL) problem, which, for a single patient can be described as follows.
Let $\mathcal{X}$ be a set of potential driver genes and $\mathcal{Y}$ be a set of expression altered (outlier) genes. Then, for a user defined k, HIT'nDRIVE can aim to return k potential driver genes as solution to the following optimization problem:

$$\arg\min_{X \subseteq \mathcal{X}, |X|=k} \max_{y \in \mathcal{Y}} H_{X,y}$$

where $H_{X,y}$ denotes the multi-hitting time from the gene set X to the gene y.
RWFL problem resembles the standard (minimax) "facility location" problem in which one seeks a set of nodes as facilities in a graph such that the maximum distance from any node in the graph to its closest facility is minimized. RWFL differs from standard facility location by its use of $H_{X,y}$ as a distance measure between a collection of nodes to any other node, which aims to capture the uncertainty in molecular interactions during the propagation of one or more signals, by random walks starting from one or more origins (reminiscent of the underlying Brownian motion). Since the standard facility location is an NP-hard problem, RWFL problem is NP hard as well. As shown in the next section, we overcome this difficulty by introducing a good estimate on the multi-hitting time that helps us to reduce RWFL problem to the weighted multi-set cover problem, which we solve through an ILP formulation in Section 3. (Although the use of set-cover for representing the most parsimonious solution in a bioinformatics context is not new [24], to the best of our knowledge this is the first use of the multi-set cover formulation for maximum parsimony.) In this formulation, we use a slightly different objective: given a user defined upper bound on the maximum multi-hitting time, we now aim to minimize the number of potential drivers that can "cover" (a user defined proportion of) the outlier genes. For more than one patient, we minimize the number of drivers that can "cover" (a user defined proportion of) patient-specific outliers such that each such outlier is covered by potential drivers that are aberrant in that patient.

2.1 Estimating Hitting Time on a Protein-Protein Interaction (PPI) Network

As mentioned before, HIT'nDRIVE estimates the multi-hitting time $H(U, v)$ between a set of nodes U and a single node v, as a function of independent hitting times $H(u, v)$ for all $u \in U$ - as will be shown later. However, even computing $H(u, v)$ is not a trivial task in a general graph $G = (V, E)$ as it requires a solution to a system of $|V|$ linear equations with $|V|$ variables. Below we show how to efficiently calculate $H(u, v)$ for all $u, v \in V$ for a graph $G = (V, E)$ with constant average degree and small diameter - as per the available human protein interaction network (or any small world network).

Let $H_{max} = \max_{u,v}\{H_{u,v}\}$. Our aim is to estimate $H_{u,v}$ empirically by performing independent random walks and taking the average of the observed hitting times. More formally, for any given number of iterations $m > 1$ and pair $u, v \in V$, let $X_1, X_2, ..., X_m$ be a sequence of independent random variables which have the same distribution as $\tau_{u,v}$ for every $1 \leq i \leq m$. Then the empirical hitting time is defined as $\tilde{H}_{u,v} = \frac{1}{m} \cdot \sum_{i=1}^{m} X_i$. The following theorem shows how fast $\tilde{H}_{u,v}$ converges to $H_{u,v}$.

Theorem 1. *Assume that G is a graph such that the maximum hitting time satisfies $H_{max} \leq Cn$ for some constant $C > 0$ and let u, v be an arbitrary pair of nodes. Then for any $\varepsilon \in [\frac{1}{n^4}, 1]$, after $m = (128C)^2(1/\varepsilon)^2(\log_2 n)^3$ iterations, the returned estimate $\tilde{H}_{u,v}$ satisfies*

$$\Pr\left[|\tilde{H}_{u,v} - H_{u,v}| \leq \varepsilon n\right] \geq 1 - n^{-3}.$$

Moreover, with probability at least $1 - n^{-7}$, the total number of random walk hops made is at most $m \cdot 32Cn \log_2 n = O((1/\varepsilon)^2 n \log^4 n)$.

We provide the proof of Theorem 1 in the Supplements. To obtain the empirical estimates of all n^2 hitting times $H_{u,v}$ efficiently, observe that taking a single random walk starting from u until all nodes are visited gives an estimate for all n hitting times $H_{u,v}$ with $v \in V$. Since for fixed $v \in V$, all m estimates for $H_{u,v}$ (coming from m iterations) are independent, we conclude by the first statement of Theorem 1 and the union bound that with probability at least $1 - n^{-2}$, for fixed $u \in V$ all n estimates $\tilde{H}_{u,v}$ approximate $H_{u,v}$ up to an additive error of εn. Similarly, the total number of random walk hops to obtain all these n approximations is $O((1/\varepsilon)^2 n \log^4 n)$ with probability at least $1 - n^{-6}$. Finally, we do the above procedure for all n possible starting vertices $u \in V$, so that with probability at least $1 - n^{-1}$, we have an εn-additive approximation for each of the n^2 hitting times, and the total number of random walk hops is $O((1/\varepsilon)^2 n^2 \log^4 n)$ with probability at least $1 - n^{-5}$.

2.2 Estimating Multi-Source Hitting Time via Single-Source Hitting Times

Given $U = \{u_1, u_2, \ldots, u_k\}$, we now show how to estimate $H_{U,v}$ by a function of independent pairwise hitting times $H_{u_i,v}$ for all $u_i \in U$. A natural estimate is

$$H_{U,v} \approx \frac{1}{\sum_{i=1}^{k} \frac{1}{H_{u_i,v}}} \tag{1}$$

Let the conductance of graph G be defined as $\Phi(G) = \min_{\emptyset \subsetneq S \subsetneq V} \frac{|E(S,V\setminus S)|}{\min\{\mathrm{vol}(S),\mathrm{vol}(V\setminus S)\}}$. Many real-world networks including preferential attachment graphs are known to have large conductance [25]. For such graphs, our next theorem provides mathematical evidence for the accuracy of our estimate in (1).

Theorem 2. *Let $G = (V, E)$ be any graph with constant conductance $\Phi > 0$. Then there is an integer $C = C(\Phi) > 0$ such that, given an integer k, a set of nodes $U = \{u_1, u_2, \ldots, u_k\}$ and node $v \in V$ satisfying $\frac{1}{k \cdot \frac{\deg(v)}{2|E|}} \geq \log^{1.5} n$, the following inequality holds:*

$$H_{U,v} \leq C \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(v)}{2|E|}}.$$

In particular, for any pair of nodes u, v with $\deg(v) \leq \frac{2|E|}{\log^{1.5} n}$ we have $H_{u,v} = O(\frac{|E|}{\deg(v)})$.

We provide the proof in the Supplements. Note that the bound in Theorem 2 differs from our estimate in equation (1) in that $\frac{1}{H_{u_i,v}}$ is replaced by $\frac{\deg(v)}{2|E|}$. However, for graphs with constant conductance, we have $H_{u,v} \leq H_{\pi,v} + O(\log n)$, where $H_{\pi,v}$ is the hitting time for a random walk starting according to the stationary distribution π, given by $\pi(w) = \frac{\deg(w)}{2|E|}$ for every $w \in V$. Hence $\frac{2|E|}{\deg(v)} = H_{v,v} \leq H_{\pi,v} + O(\log n)$. Since $H_{\pi,v} = \sum_{u \in U} \pi(u) \cdot H_{u,v}$, it follows that, given any fixed node v, it holds for "most nodes" u that $H_{u,v}$ is not much smaller than $\frac{2|E|}{\deg(v)} - O(\log n)$.

3 Reformulation of RWFL as a Weighted Multi-Set Cover Problem

Since RWFL is NP-hard we reduce it to the weighted set cover problem, which we solve via an ILP formulation. This formulation also generalizes RWFL to allow patient-specific drivers and outlier genes. Consider a bipartite graph $G_{bip}(\mathcal{X}, \mathcal{Y}, \mathcal{E})$ where $\mathcal{X}$ is the set of aberrant genes, $\mathcal{Y}$ is the set of patient-specific expression altered genes, and $\mathcal{E}$ is the set of edges. If gene g_i is mutated in a patient p, we set edges between g_i and all of the expression altered genes in the same patient (g_j, p) where the edges are weighted by the inverse pairwise-hitting times $w_{i,j} := H^{-1}_{g_i,g_j}$; see the Figure 1 for more details.

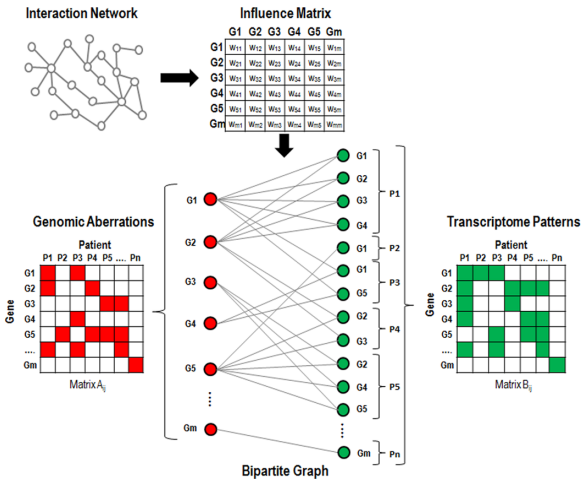

Fig. 1. Schematic overview of construction of bipartite graph in HIT'nDRIVE. The influence matrix derived from the interaction network contains the *inverse* hitting time between every pair of genes. A and B are gene-patient matrices showing the genomic abberations and expression alteration events, respectively. The red color in A indicates the aberration status of a gene in a patient. Similarly, the green color in B indicate expression altered genes in a patient. The edges in the bipartite graph are weighted by the inverse hitting time within the PPI network.

We now define a minimum weighted multi-set cover (WMSC) problem on G_{bip}, whose solution provides an exact solution to RWFL problem, provided our estimate of the multi-hitting times are accurate, i.e.

$$\arg\min_{X\subseteq\mathcal{X}}|X| \quad \text{such that} \quad \max_{y\in\mathcal{Y}} H_{X,y} \leq \Delta \tag{2}$$

where Δ is the maximum allowed multi-hitting time from the drivers to any expression altered gene.

WMSC asks to compute as the potential driver gene set, the smallest set which "sufficiently" covers "most" of the patient specific expression altered genes:

$$\arg\min_{X\in\mathcal{X}} \min_{Y\subseteq\mathcal{Y},|Y|\geq\alpha|\mathcal{Y}|} |X| \quad \text{such that} \quad \forall y\in Y: \sum_{x\in X} w_{x,y} \geq \gamma_y \tag{3}$$

where $0 < \alpha \leq 1$ represents the fraction of patient-specific expression altered genes that we believe are causally linked to the potential drivers. The left-hand-side of the constraints in (2) and (3) are related by $H^{-1}_{X,y} \approx \sum_{x\in X} H^{-1}_{x,y}$, as mentioned in Section 2.2. The introduction of γ_y makes it possible to control the minimum amount of "coverage" needed for *individual* expression alteration events (each patient potentially indicates a unique expression alteration event for each gene).

3.1 An ILP Formulation for WMSC

We formulate WMSC as an ILP and solve it using an off-the-shelf ILP solver. The ILP formulation for our combinatorial optimization problem is as Figure 2 where there is a binary variable x_i, y_j, e_{ij}, respectively, for each potential driver, expression alteration event, and edge in the bipartite graph. The first constraint ensures that a selected driver contributes to the coverage of each of the expression alteration events it is connected to - in each patient. The second constraint ensures that selected (patient-specific) driver genes cover at least a (γ) fraction of the sum of all incoming edge weights to each expression alteration event. This constraint corresponds to setting a lower bound on the joint influence (i.e. our estimate on the inverse of multi-hitting time) of selected (patient specific) drivers on an expression alteration event. The third constraint ensures that the selected driver genes collectively cover at least an α fraction of the set of expression alteration events.

$$
\begin{array}{l}
\min_{x_1,\ldots,x_{|\mathcal{X}|}} \sum_i x_i \\
\text{s.t.} \\
\forall i, j : x_i = e_{ij} \\
\forall j : \sum_i e_{ij} w_{ij} \geq y_j \gamma \sum_i w_{ij} \\
\sum_j y_j \geq \alpha |\mathcal{Y}| \\
x_i, e_{ij}, y_j \in \{0, 1\}
\end{array}
$$

Fig. 2. ILP formulation.

4 Evaluation Framework

Evaluating computational methods for predicting cancer drivers is challenging in the absence of the ground truth (i.e. follow-up biological experiments). We refer to previous studies [18] that observe the overlap between predicted driver genes and known cancer genes compiled in public resources such as the Cancer Gene Census (CGC) database [26] or the Catalogue of Somatic Mutations in Cancer (COSMIC) database [27] and we provide those numbers as well. However, we mainly focused on testing whether our predictions provide insight into the cancer phenotype and improve classification accuracy on an independent cancer dataset. The classifiers we evaluate are based on network "modules", a set of functionally related genes (e.g. in a signaling pathway), which are connected in an interaction network and include at least one potential driver. They then use module features, such as the average expression of genes in the module, for phenotype classification. Using such module features, we hope that the classifier in use does not *overfit* on rare drivers and is able to *generalize* the signal coming from rare drivers to new patients.

For classification purposes we primarily use OptDis [22] for *de novo* identification of modules which include (i.e. are seeded by) at least one predicted driver gene. In general, OptDis performs supervised dimensionality reduction on the set of connected subnetworks. It projects the high dimensional space of all connected subnetworks to a user-specified lower dimensional space of subnetworks such that, in the new space, the samples belonging to the same (different) class are closer (respectively, more distant) to each other with respect to a normalized distance

measure (typically L_1). Since the human PPI network has a small diameter, there is significant overlap between many modules seeded by potential driver genes. In order to limit the number of overlapping modules (and achieve further dimensionality reduction) we first compute the top 10 modules seeded by each driver gene that have the best individual "discriminative scores" (a linear combination of the average in-class distance and out-class distance [22]). The modules seeded by all potential drivers are then collectively sorted based on their discriminative score. Among these modules, we greedily pick a subset in a way that the i^{th} module is added to our result subset R if its maximum pairwise node overlap with any module already in R is no more than a user-defined threshold.

5 Experiments

We use a publicly available cancer dataset representing matched genomic aberration (somatic mutation, copy-number aberration) and transcriptomic patterns (gene-expression data) of 156 Glioblastoma Multiforme (GBM) samples [5] from The Cancer Genome Atlas (TCGA). We make use of a global network of protein-protein interaction (PPI) from the Human Protein Reference Database (HPRD) version April 2010 [28] to derive the influence values based on the hitting time. We use the same PPI Network for module identification using our modification to OptDis. We ran HIT'nDRIVE with different combinations of values for the variables α and γ as given in Figure 3-A. For a fixed γ, the number of selected driver genes increased linearly with the value of α. The increase in the number of drivers is expected as more drivers are required to cover larger fraction of abnormal expression events.

Evaluation Based on CGC and COSMIC Databases. To assess whether the genes identified by HIT'nDRIVE are essential players in cancer, we first analyzed the concordance of the predicted drivers with the genes annotated in CGC and COSMIC database. Gene sets resulting with the parameters $\gamma = 0.7$ and $\alpha = \{0.1, 0.2, ..., 0.9\}$ were analyzed (Figure 3-B). The fraction of driver genes affiliated with cancer in the CGC and COSMIC databases increase with increasing values of α.The remainder of results are obtained for parameter values $\gamma = 0.7$ and $\alpha = 0.9$ this results in 107 driver genes covering the majority (22933) of outlier genes in 156 patients.

Phenotype Classification Using Dysregulated Modules Seeded with the Predicted Drivers. We evaluated the driver genes identified by HIT'nDRIVE using phenotype classification (as described in Section 4 and results are shown in Figure 4). Briefly, drivers identified from the TCGA dataset were used as seeds for discovering discriminative subnetwork modules. The module expression profiles

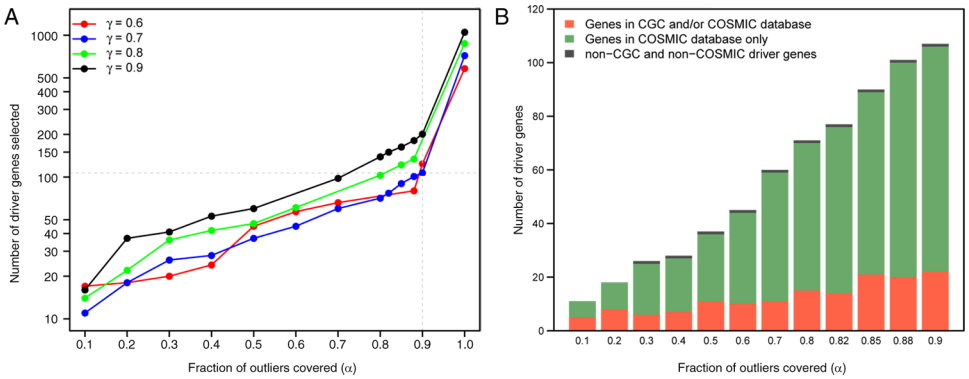

Fig. 3. Behavior of HIT'nDRIVE as a function of α and γ. (A) The number of selected drivers and covered outliers as α increases for various values of γ. Note that some of the data points are missing for the problems which could not be solved within 48 hours. (B) Concordance of GBM driver genes with that of COSMIC and Cancer Gene Census database for $\gamma = 0.7$.

were used to classify normal vs. glioblastoma samples through repeated cross-validation on the validation dataset. First, HIT'nDRIVE using hitting time based influence values, was compared against DriverNet, which greedily identifies driver genes using direct gene interactions from the HPRD network. Across the appreciable range of discriminative modules discovered by OptDis, HIT'nDRIVE demonstrates better accuracy in classifying the cancer phenotype, with a maximum accuracy of 97.05% and a mean accuracy of 94.52% (Figure 4). Next, comparing the HIT'nDRIVE deduced drivers against a comparable number of genes with the highest node-degrees in the PPI network reveals a clear advantage to HITnDRIVE. This trend was observed when genes were used as individual classification features (blue vs. orange plots) as well as when they were used as seeds for module-based features (red vs. brown plots). Comparing the classification accuracy of HITnDRIVE deduced drivers against 107 genes randomly selected from the entire list of aberrant genes (red vs. grey plots) provides additional support for the relevance of drivers selected by HITnDRIVE. This is also confirmed by comparing the performance of hitting-time based influence values against those derived from the diffusion model [16] (red vs. black plots) both employed by HITnDRIVE.

Sensitivity of HIT'nDRIVE to Small Perturbations of the PPI Network. We perturbed the PPI network by swapping endpoints at random of 20% edges and recalculated pairwise hitting times. We observed that almost all changes are less than 10% relative to the original values, most of them being between 1% and 5%. However, impact on accuracy of classification using HIT'nDRIVE output can be noticed in Figure 4.

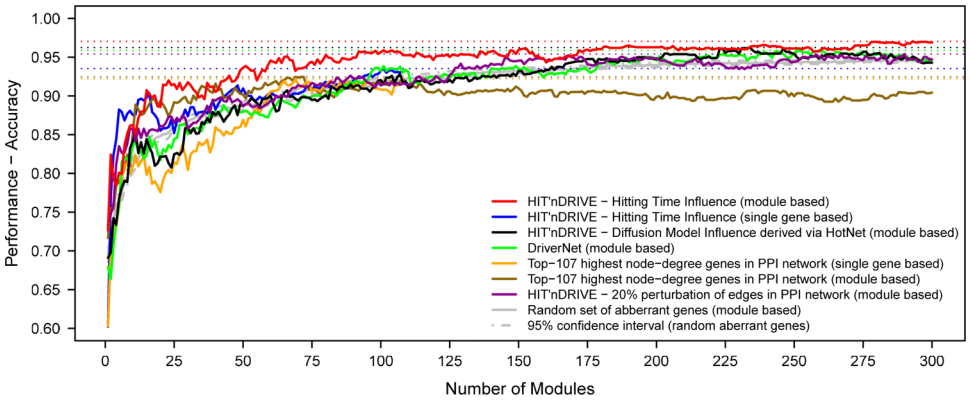

Fig. 4. Phenotype classification using the identified drivers obtained by various methods. The dysregulated sets of modules seeded by the 107 chosen drivers are used to predict phenotype in the validation dataset using using k-nearest neighbour classifier with k=1. We used the HPRD-PPI Network for module identification using our modification to OptDis.

Prediction of Frequent and Rare Drivers. The 107 driver genes nominated by HIT'nDRIVE are aberrated at varying frequencies in the tumor population (Figure 5-A). CHEK2 and EGFR are the two most frequently aberrated drivers (at 46.8% and 42.3% respectively), followed by CDKN2A (31.4%), MTAP (30.1%) and CDKN2B (29.5%). Some of these frequent drivers harbour different types of genomic aberrations in different patients. For example, EGFR shows somatic mutation and high copy-number gain in 14.2% and 32.7% of the patients, respectively. Similarly, PTEN harbours somatic mutation in 12.8% and homozygous deletion in 3.9% of the patients. Amplification in EGFR, PDGFRA, mutations in CHEK2, TP53, PTEN, RB1, and deletions in CDKN2A have been previously associated with GBM [5, 29, 30]. HIT'nDRIVE also identified infrequent drivers, which we defined as genes that are genomically aberrant in at most 2% of the cases. Out of 26 (16.66%) rare driver genes identified, four genes (MYST4, FLI1, BMPR1A and BRCA2) were implicated in the CGC database. Despite being aberrant in a small fraction of patients, the rare drivers are specifically associated with cancer development, DNA repair, cell growth and migration, cell death and survival. Some rare drivers like MAG and BMPR1A have also been closely linked with GBM progression [31, 32].

Prediction of Low-degree and High-degree Drivers. The drivers predicted by HIT'nDRIVE include a number of well-known high-degree "hubs" such as TP53, EGFR, RB1 and BRCA1, which occupy the central position (with high degree and high betweenness, i.e. the proportion of shortest paths between all pairs of nodes that go through that node, and high degree - computed by the igraph [33] R package.) in the PPI network (Figure 5-B). If these genes are perturbed, they

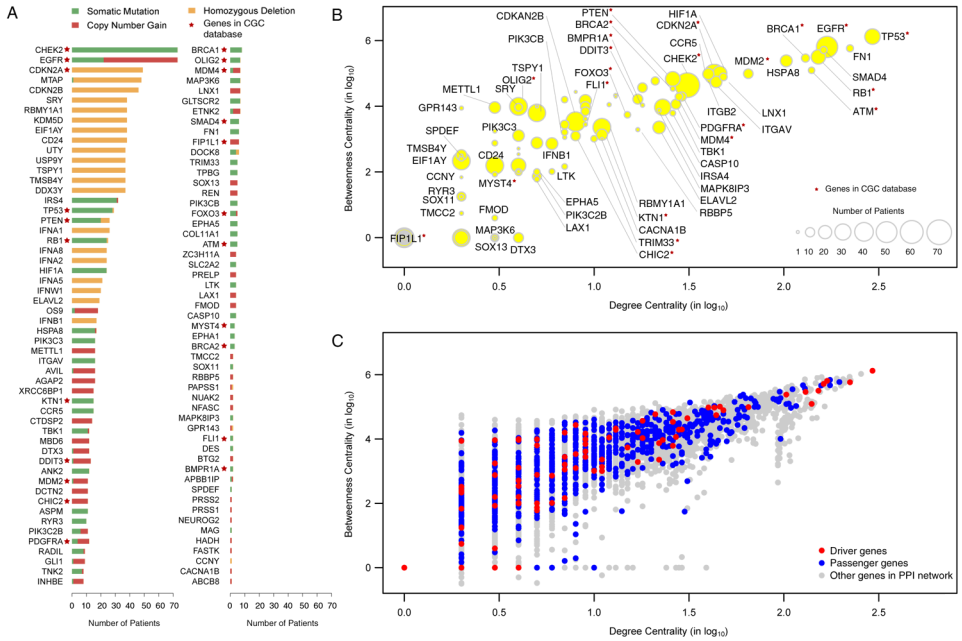

Fig. 5. Characteristics of driver genes of GBM predicted by HIT'nDRIVE. (A) Recurrence frequency of the aberration in the driver genes predicted by HIT'nDRIVE. (B) The centrality of the predicted drivers in the PPI network.The size of the circles is proportional to the recurrence frequency of the genomic aberration of the gene. (C) Centrality of the "driver" and "passenger" genes is colored by red and blue dots respectively; all other nodes in the PPI network apart from the driver and passenger genes are represented as grey dots.

dysregulate several other genes and the associated signaling pathways. Moreover, HIT'nDRIVE also identified low-degree genes (such as FIP1L1, SOX11 and RYR3) that reside in the periphery of the PPI network. Some of these low-degree genes are only aberrant in a small fraction of patients. Since driver genes and passenger genes display similar network characteristics (Figure 5-C), and identified driver genes have both low and high degrees in the network, HIT'nDRIVE likely selects drivers irrespective of known network biases.

6 Conclusion and Future Work

We have presented HIT'nDRIVE, a combinatorial method to capture the collective effects of driver gene aberrations on possibly distant "outlier" genes based on what we call the "random-walk facility location" (RWFL) problem. We introduced the notion of "multi-source hitting time" and presented efficient and accurate methods to estimate it based on single-source hitting time in large-scale networks. We applied HIT'nDRIVE to identify genes subject to somatic mutation and copy number in GBM. Our results showed that the predicted driver genes identified by

HIT'nDRIVE are well-supported in databases of important cancer genes. Furthermore, these drivers were able to perform phenotype predictions more accurately than the alternative approaches. Importantly, the discovery of these drivers were not biased by the frequency of aberration and/or the degree of a gene in the PPI network. Our approach can easily integrate various aberration types such as single nucleotide changes, copy number changes, structural variations, and splice variations. Furthermore, it can be straightforwardly extended to incorporate epigenome and/or gene-fusions data. As gene networks increase in density and volume of interaction, HIT'nDRIVE will be able to capture such improvements naturally. Finally our method is well suited to identify patient-specific driver-aberrations which can potentially be used as therapeutic targets.

Supplements: All supplementary material can be found and downloaded at http://compbio.cs.sfu.ca/software-hitndrive

References

1. Stratton, M.R., Campbell, P.J., Futreal, P.A.: The cancer genome. Nature **458**(7239) (April 2009) 719–24
2. Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., et al.: Patterns of somatic mutation in human cancer genomes. Nature **446**(7132) (March 2007) 153–8
3. Greenman, C., Wooster, R., Futreal, P.A., Stratton, M.R., Easton, D.F.: Statistical analysis of pathogenicity of somatic mutations in cancer. Genetics **173**(4) (August 2006) 2187–98
4. Youn, A., Simon, R.: Identifying cancer driver genes in tumor genome sequencing studies. Bioinformatics (Oxford, England) **27**(2) (January 2011) 175–81
5. Parsons, D.W., Jones, S., Zhang, X., Lin, J.C.H., Leary, R.J., Angenendt, P., et al.: An integrated genomic analysis of human glioblastoma multiforme. Science (New York, N.Y.) **321**(5897) (September 2008) 1807–12
6. Cancer Genome Atlas Network: Integrated genomic analyses of ovarian carcinoma. Nature **474**(7353) (June 2011) 609–15
7. Cancer Genome Atlas Network: Comprehensive molecular characterization of human colon and rectal cancer. Nature **487**(7407) (July 2012) 330–7
8. Greaves, M., Maley, C.C.: Clonal evolution in cancer. Nature **481**(7381) (January 2012) 306–13
9. Ding, L., Ley, T.J., Larson, D.E., Miller, C.a., Koboldt, D.C., et al.: Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing. Nature **481**(7382) (January 2012) 506–10
10. Akavia, U.D., Litvin, O., Kim, J., Sanchez-Garcia, F., Kotliar, D., et al.: An integrated approach to uncover drivers of cancer. Cell **143**(6) (December 2010) 1005–17
11. Masica, D.L., Karchin, R.: Correlation of somatic mutation and expression identifies genes important in human glioblastoma progression and survival. Cancer research **71**(13) (July 2011) 4550–61
12. Leiserson, M.D.M., Blokh, D., Sharan, R., Raphael, B.J.: Simultaneous identification of multiple driver pathways in cancer. PLoS computational biology **9**(5) (May 2013) e1003054
13. Ciriello, G., Cerami, E., Sander, C., Schultz, N.: Mutual exclusivity analysis identifies oncogenic network modules. Genome research **22**(2) (February 2012) 398–406
14. Kim, Y.A., Wuchty, S., Przytycka, T.M.: Identifying causal genes and dysregulated pathways in complex diseases. PLoS computational biology **7**(3) (March 2011) e1001095

15. Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., et al.: Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics (Oxford, England) **26**(12) (June 2010) i237–45
16. Vandin, F., Upfal, E., Raphael, B.J.: Algorithms for detecting significantly mutated pathways in cancer. Journal of computational biology : a journal of computational molecular cell biology **18**(3) (March 2011) 507–22
17. Paull, E.O., Carlin, D.E., Niepel, M., Sorger, P.K., Haussler, D., et al.: Discovering causal pathways linking genomic events to transcriptional states using Tied Diffusion Through Interacting Events (TieDIE). Bioinformatics (Oxford, England) (September 2013) 1–8
18. Bashashati, A., Haffari, G., Ding, J., Ha, G., Lui, K., et al.: DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. Genome biology **13**(12) (December 2012) R124
19. Liben-Nowell, D., Kleinberg, J.: The link-prediction problem for social networks. Journal of the American Society for Information Science and Technology **58**(7) (May 2007) 1019–1031
20. Hopcroft, J., Sheldon, D.: Manipulation-resistant reputations using hitting time. In: Algorithms and Models for the Web-Graph. Springer (2007) 68–81
21. Tetali, P.: Design of on-line algorithms using hitting times. SIAM J. Comput. **28**(4) (1999) 1232–1246
22. Dao, P., Wang, K., Collins, C., Ester, M., Lapuk, A., Sahinalp, S.C.: Optimally discriminative subnetwork markers predict response to chemotherapy. Bioinformatics **27**(13) (Jul 2011)
23. Levin, D.A., Peres, Y., Wilmer, E.L.: Markov Chains and Mixing Times. American Mathematical Society (2008)
24. Hormozdiari, F., Alkan, C., Eichler, E.E., Sahinalp, S.C.: Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. Genome research **19**(7) (July 2009) 1270–1278
25. Mihail, M., Papadimitriou, C.H., Saberi, A.: On certain connectivity properties of the internet topology. J. Comput. Syst. Sci. **72**(2) (2006) 239–251
26. Futreal, P.A., Coin, L., Marshall, M., Down, T., Hubbard, T., et al.: A census of human cancer genes. Nature reviews. Cancer **4**(3) (March 2004) 177–83
27. Forbes, S.a., Bindal, N., Bamford, S., Cole, C., Kok, C.Y., et al.: COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research **39**(Database issue) (January 2011) D945–50
28. Prasad, T.S.K., Kandasamy, K., Pandey, A.: Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods in molecular biology (Clifton, N.J.) **577** (January 2009) 67–79
29. Cancer Genome Atlas Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature **455**(7216) (October 2008) 1061–8
30. Verhaak, R.G.W., Hoadley, K.a., Purdom, E., Wang, V., Qi, Y., et al.: Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. Cancer cell **17**(1) (January 2010) 98–110
31. McKerracher, L., David, S., Jackson, D.L., Kottis, V., Dunn, R.J., et al.: Identification of myelin-associated glycoprotein as a major myelin-derived inhibitor of neurite growth. Neuron **13**(4) (October 1994) 805–11
32. Piccirillo, S.G.M., Reynolds, B.a., Zanetti, N., Lamorte, G., Binda, E., et al.: Bone morphogenetic proteins inhibit the tumorigenic potential of human brain tumour-initiating cells. Nature **444**(7120) (December 2006) 761–5
33. Csardi, G., Nepusz, T.: The igraph software package for complex network research. InterJournal **Complex Systems** (2006) 1695

Appendix

Datasets For publically available Glioblastoma multiforme (GBM) [5], somatic mutations (level 2), array based copy-number (level 1) and microarray gene-expression data (level 3) were obtained from TCGA data portal. GBM represented 156 samples which had matched somatic mutation, copy-number and gene-expression data available. Two gene-expression datasets, GSE11882 and GSE7696, were obtained from the Gene Expression Omnibus (GEO) repository and merged after z-score transformation to obtain a validation dataset containing the gene-expression profile of 177 normal brain tissues and 80 GBM samples.

Somatic mutation Somatic mutation calls for each participant (level 2) data generated from Illumina Genome Analyzer DNA Sequencing were analyzed. Only missense mutation, nonsense mutation and splice-site SNPs were marked as a somatic-mutation aberrant event.

Copy number aberration Array based copy-number (level 1) data files were downloaded via the TCGA data portal. These Agilent FE format sample files were loaded into BioDiscovery Nexus Copy Number software v7.0, where quality was assessed and data was visualized and analyzed. All samples were mapped to the most recent genome build (hg 19, NCBI build 37) via Agilent probe identifiers and annotation (downloaded from Agilents website) based on the 1M SurePrint G3 Human CGH Microarray 1x1M design platform. BioDiscoverys FASST2 Segmentation Algorithm, a Hidden Markov Model (HMM) based approach, was used to make copy number calls. The FASST2 algorithm, unlike other common HMM methods for copy number estimation, does not aim to estimate the copy number state at each probe but uses many states to cover more possibilities, such as mosaic events. These state values are then used to make calls based on a log-ratio threshold. The significance threshold for segmentation was set at 5.0E-6 also requiring a minimum of 3 probes per segment and a maximum probe spacing of 1000 between adjacent probes before breaking a segment. The log ratio thresholds for single copy gain and single copy loss were set at 0.2 and -0.23, respectively. The log ratio thresholds for two or more copy gain and homozygous loss were set at 1.14 and -1.1 respectively. Upon loading of raw data files, signal intensities are normalized via division by mean. All samples are corrected for GC wave content using a systematic correction algorithm. Only the high confidence copy number aberrations i.e. high copy number gain or homozygous deletion were marked as a copy-number aberrant event. Finally, the genes that harbour either a somatic-mutation aberrant event or a copy-number aberrant event were taken to be the final list of abberant genes at the genomic level.

Derivation of outlier matrix For GBM, microarray gene-expression data (level-3) from Affymetrix HT Human Genome U133 Array Plate Set were analyzed. The outlier genes are defined as those values that are outside the 2.7 standard deviation range of the expression values of the gene across all the patients.

Cross Validation for Glioblastoma The classification performance of dysregulated module seeded by driver gene was evaluated through 5-fold cross-validation repeated five times on the validation dataset. A weighted *k*-nearest-neighbor (knn) classifier with k=1 which was originally developed for small molecule classification was used

Evaluating our multi-hitting time estimate. For the purpose of evaluating our multi-hitting time estimate in the human PPI network, we picked the following 10 genes at random: ATP7A, BMP15, CNPY2, FHL1, PDZD4, PIK3CG, RAB3D, TRIM3, TSPY1, ZRSR2. On this set, we computed the exact solution to the RWFL problem for 3 "facilities" using a brute-force approach: CNPY2, RAB3D and TRIM3. Then we applied Hit'nDrive where only the above 10 genes were kept on the left-hand side of the bipartite graph, using parameters $\alpha = 0.99$ and $\gamma = 0.7$. The solution obtained included the three genes offered by the exact solution, i.e. CNPY2, RAB3D, TRIM3, plus one more gene, TSPY1, suggesting that our estimate of multi-hitting time works well in practice.

Functional and Pathway Enrichment of the Driver Genes. Ingenuity Pathway Analysis (IPA; Ingenuity Systems, www.ingenuity.com) software was used to find significant functional and pathway enrichment of the predicted driver genes. An enriched function or pathway with Benjamini-Hochberg corrected p-value <= 0 : 00001 was selected as significant.
The collective set of 107 driver genes are significantly enriched for relevant biological functions such as neurological disease, cancer, cellular growth, cell death and survival (Figure Appendix-1-A). Additionally, these driver genes were found to modulate various oncogeneic and tumor supressor pathways such as p53 signalling (Figure Appendix-1-B) further validating our method.

1.1 Proof of Theorem 1

Proof. By Markov's inequality, we have for every $1 \leq i \leq m$, $\Pr[X_i \geq 2H_{max}] \leq \frac{1}{2}$. Diving the random walk into k consecutive sections of length $2H_{max}$ yields for any

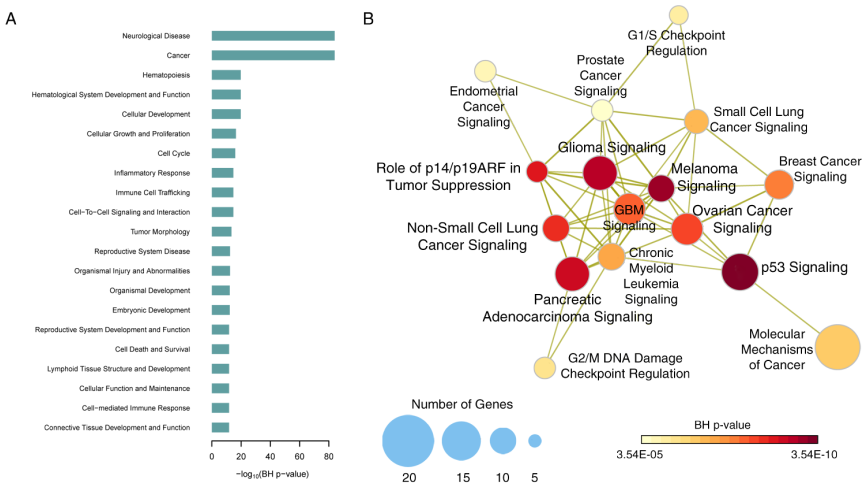

Fig. Appendix-1. Functional and pathway enrichment of the driver genes (A) Functions enriched within the driver genes selected by HIT'nDRIVE. (B) Network displaying the pathway enrichment of the selected drivers in GBM cohort. The size of the node is proportional to the number of genes enriched for the pathway. Heat color has been assigned to the node according to the significance of the pathway or function. The relationship in the network is in accordance to the correlation between pathways. Correlation between the pathway is drived using the driver genesets enriched within each of the significant pathway. All of the functions and pathways are significant with Benjamini-Hochberg corrected p-value <= 0.00001.

integer $k \geq 1$,

$$\Pr\left[X_i \geq k \cdot 2H_{max}\right] \leq \left(\frac{1}{2}\right)^k.$$

Let us define $\mathcal{E}$ as the event which occurs if for every $1 \leq i \leq m$, $X_i \leq 32 \log_2 n \cdot H_{max}$. By the union bound,

$$\Pr\left[\mathcal{E}\right] \geq 1 - m \cdot \left(\frac{1}{2}\right)^{16 \log_2 n} = 1 - m \cdot n^{-16} \geq 1 - n^{-7},$$

where the last inequality is due to the definition of m and the lower bound on ε. Observe that if the event $\mathcal{E}$ occurs, then the total number of random walk steps made is at most $m \cdot 32 \log_2 n \cdot H_{max} \leq m \cdot 32Cn \log_2 n$, which yields the second statement of the theorem.

We now prove the first statement of the theorem. Conditioning the expectation of X_i yields

$$E\left[X_i\right] = \Pr\left[\mathcal{E}\right] \cdot E\left[X_i \mid \mathcal{E}\right] + \Pr\left[\neg\mathcal{E}\right] \cdot E\left[X_i \mid \neg\mathcal{E}\right].$$

By the memoryless property of the random walk,

$$E[X_i \mid \neg\mathcal{E}] \leq 32Cn\log_2 n + H_{max}.$$

Consequently,

$$E[X_i] \leq 1 \cdot E[X_i \mid \mathcal{E}] + n^{-7} \cdot (32Cn\log_2 n + Cn) \leq E[X_i \mid \mathcal{E}] + \frac{\varepsilon}{2} \cdot n.$$

By definition of $\mathcal{E}$, $E[X_i] \geq E[X_i \mid \mathcal{E}]$, and combining the previous two inequalities yields

$$|E[X_i] - E[X_i \mid \mathcal{E}]| \leq \frac{\varepsilon}{2} \cdot n. \tag{4}$$

Note that in the probability space conditional on the event $\mathcal{E}$, the random variables $X_1, X_2, ..., X_m$ are mutually independent, identically distributed random variables with expectation $E[X_1 \mid \mathcal{E}]$ each. Furthermore, each random variable takes values in $\{1, 2, ..., 32Cn\log_2 n\}$. Hence Hoeffding's inequality gives for any $\lambda > 0$,

$$\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \geq \lambda \,\middle|\, \mathcal{E}\right] \leq 2 \cdot \exp\left(-\frac{2\lambda^2}{m \cdot (32Cn\log_2 n)^2}\right).$$

Choosing $\lambda = 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5}$ yields

$$\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \geq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5} \,\middle|\, \mathcal{E}\right] \leq 2n^{-4}.$$

With our lower bound on $\Pr[\mathcal{E}]$, we conclude that

$$\begin{aligned}
&\Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \leq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5}\right] \\
&\geq \Pr[\mathcal{E}] \cdot \Pr\left[\left|\sum_{i=1}^{m} X_i - m \cdot E[X_1 \mid \mathcal{E}]\right| \leq 64C\sqrt{m} \cdot n \cdot (\log_2 n)^{1.5} \,\middle|\, \mathcal{E}\right] \geq (1 - n^{-7}) \cdot (1 - 2n^{-4}) \geq 1 - n^{-3}.
\end{aligned}$$

If the above event occurs, then our returned estimate $\tilde{H}_{u,v}$ satisfies

$$\left|\tilde{H}_{u,v} - E[X_1 \mid \mathcal{E}]\right| < \frac{64C(\log_2 n)^{1.5}}{\sqrt{m}} \cdot n = \frac{\varepsilon}{2} \cdot n,$$

where the last equality follows from the definition of m. Combining this with equation (4) yields

$$\left|\tilde{H}_{u,v} - E[X_1]\right| \leq \left|\tilde{H}_{u,v} - E[X_1 \mid \mathcal{E}]\right| + |E[X_1 \mid \mathcal{E}] - E[X_1]| = 2 \cdot \frac{\varepsilon}{2} \cdot n = \varepsilon \cdot n,$$

which completes the proof of the first statement as $\mathrm{E}[X_1] = H_{u,v}$. ■

1.2 Proof of Theorem 2

For the proof of Theorem 2, it will be convenient to consider a lazy version of the random walk which stays at the current node in each step with probability $1/2$. Note that any hitting time (single-source or multi-source) of the lazy verion of the random walk is always an upper bound on the corresponding hitting time of the standard random walk.

Lemma 3. *Let $G = (V, E)$ be a graph with constant conductance $\Phi > 0$. For any pair of nodes $u, v \in V$ and number of steps t with $\omega(\log n) \leq t \leq \frac{2|E|}{\deg(v)}$, let $\mathcal{A}_{u,v,t}$ be the event that a random walk starting from u visits v within t steps. Then*

$$\Pr[\,\mathcal{A}_{u,v,t}\,] \geq \frac{\Phi^2}{280} \cdot t \cdot \frac{\deg(v)}{2|E|}.$$

Proof. We first record the following useful inequality (cf. [23]). Let $P^s_{x,y}$ be the probability that a random walk starting at x visits node y in step s. Then,

$$\left| P^s_{x,y} - \frac{\deg(y)}{2|E|} \right| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \cdot \lambda_{\max}^t,$$

where $\pi(w) = \frac{\deg(w)}{2|E|}$ for any $w \in V$, $\lambda_{\max} = \max\{\lambda_2, |\lambda_n|\}$ with $1 = \lambda_1 \geq \cdots \geq \lambda_n > -1$ being the eigenvalues of the transition matrix P. Since the random walk has loop probability $1/2$, $\lambda_n \geq 0$ and thus $\lambda_{\max} = \lambda_2$. Furthermore, by Cheeger's inequality, $\lambda_2 \leq 1 - \frac{\Phi^2}{8}$. Hence

$$\left| P^s_{x,y} - \frac{\deg(y)}{2|E|} \right| \leq \sqrt{\frac{\pi(y)}{\pi(x)}} \cdot \left(1 - \frac{\Phi^2}{8}\right)^t,$$

which implies for every s with $t/2 \leq s \leq t$, as $t = \omega(\log n)$,

$$\left| P^s_{u,v} - \frac{\deg(v)}{2|E|} \right| \leq n^{-4}.$$

Let X be the random variable counting the number of visits to v within the time-interval $[t/2, t]$. Then, from the above,

$$\frac{t}{2} \cdot \frac{\deg(v)}{2|E|} \leq \mathrm{E}[\,X\,] \leq 2t \cdot \frac{\deg(v)}{2|E|}.$$

To apply the second moment method, we will now analyze the variance of X, denoted by $\mathrm{V}[\,X\,]$. Note that $X = \sum_{s=t/2}^{t} X_s$, where $X_s = 1$ if the random walk visits

u in step s and $X_s = 0$ otherwise. Then,

$$\begin{aligned}
\mathrm{V}[X] &\leq \sum_{s=t/2}^{t} \mathrm{E}[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1 \wedge X_{s'} = 1] - \Pr[X_s = 1] \cdot \Pr[X_{s'} = 1] \\
&= \sum_{s=t/2}^{t} \mathrm{E}[X_s] + 2 \sum_{t/2 \leq s < s' \leq t} \Pr[X_s = 1] \cdot (\Pr[X_{s'} = 1 \mid X_s = 1] - \Pr[X_{s'} = 1]) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s < s' \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\left(\frac{\deg(v)}{2|E|} + (1 - \frac{\Phi^2}{8})^{s'-s} \right) - \left(\frac{\deg(v)}{2|E|} - n^{-4} \right) \right) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s \leq t} \sum_{1 \leq i \leq t/2} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left((1 - \frac{\Phi^2}{8})^{i} + n^{-4} \right) \\
&\leq \mathrm{E}[X] + 2 \sum_{t/2 \leq s \leq t} \left(\frac{\deg(v)}{2|E|} + n^{-4} \right) \cdot \left(\frac{8}{\Phi^2} + t/2 \cdot n^{-4} \right) \\
&\leq \mathrm{E}[X] \cdot \left(2 + \frac{32}{\Phi^2} \right) + O(n^{-2}) \leq \frac{35}{\Phi^2} \cdot \mathrm{E}[X] .
\end{aligned}$$

By the Paley-Zygmund inequality, for any $0 < \delta < 1$,

$$\Pr[X \geq \delta \cdot \mathrm{E}[X]] \geq (1-\delta)^2 \cdot \frac{\mathrm{E}[X]^2}{\mathrm{V}[X] + \mathrm{E}[X]^2} \geq (1-\delta)^2 \cdot \frac{1}{\frac{35}{\Phi^2} \cdot \frac{1}{\mathrm{E}[X]} + 1} \geq (1-\delta)^2 \cdot \frac{\Phi^2}{2 \cdot 35} \cdot \mathrm{E}[X],$$

where the last inequality follows from $\mathrm{E}[X] \leq 2$ which holds thanks to our upper bound on t. Choosing $\delta = \frac{1}{2}$ implies, as X is an integer-valued random variable,

$$\Pr[A_{u,v,t}] = \Pr[X \geq 1] \geq \Pr\left[X \geq \frac{1}{2} \cdot \mathrm{E}[X] \right] \geq \frac{\Phi^2}{8 \cdot 35} \cdot \mathrm{E}[X],$$

and due to the lower bound on $\mathrm{E}[X]$ derived earlier, the proof is finished. ∎

With the lemma at hand, we are now able to complete the proof of Theorem 2.

Proof. For any integer $\alpha \geq 1$, define $\tau = \tau(\alpha) := \alpha \cdot \frac{280}{\Phi^2} \cdot \frac{1}{\sum_{i=1}^{k} \frac{\deg(u)}{2|E|}}$. For any $1 \leq i \leq k$, let $\mathcal{E}_i$ be the event that the random walk starting from v_i does *not* visit u within τ steps. By partitioning the τ steps into consecutive sections of length $\log^{1.5} n$ and applying Lemma 3 to every section, we conclude that

$$\Pr[\mathcal{E}_i] \leq \left(1 - \frac{\Phi^2}{280} \cdot \log^{1.5} n \cdot \frac{\deg(u)}{2|E|} \right)^{\tau / \log^{1.5} n} \leq \exp\left(-\tau \cdot \frac{\Phi^2}{280} \cdot \frac{\deg(u)}{2|E|} \right).$$

As all k random walks are independent, it follows that

$$\Pr\left[\bigwedge_{i=1}^{k}\mathcal{E}_i\right] = \prod_{i=1}^{k}\Pr\left[\,\mathcal{E}_i\,\right] \le \exp\left(-\tau\cdot\sum_{i=1}^{k}\frac{\Phi^2}{280}\cdot\frac{\deg(u)}{2|E|}\right) = \exp(-\alpha) \le 2^{-\alpha}.$$

Hence the expected multi-source hitting time can be estimated as follows,

$$H_{\{v_1,\ldots,v_k\},u} \le \frac{280}{\Phi^2}\cdot\frac{1}{\sum_{i=1}^{k}\frac{\deg(u)}{2|E|}}\cdot\sum_{\alpha=1}^{\infty}\alpha\cdot 2^{-\alpha} \le \frac{560}{\Phi^2}\cdot\frac{1}{\sum_{i=1}^{k}\frac{\deg(u)}{2|E|}}$$

■