Algorithms for Detecting Significantly Mutated Pathways in Cancer

Fabio Vandin[1,⋆], Eli Upfal[2], and Benjamin J. Raphael[2,3,⋆⋆]

[1] Dipartimento di Ingegneria dell'Informazione, Università di Padova, Padova, Italy.
vandinfa@dei.unipd.it
[2] Department of Computer Science, Brown University, Providence, RI.
{eli,braphael}@cs.brown.edu
[3] Center for Computational Molecular Biology, Brown University, Providence, RI.

Abstract. Recent genome sequencing studies have shown that the somatic mutations that drive cancer development are distributed across a large number of genes. This mutational heterogeneity complicates efforts to distinguish functional mutations from sporadic, passenger mutations. Since cancer mutations are hypothesized to target a relatively small number of cellular signaling and regulatory pathways, a common approach is to assess whether known pathways are enriched for mutated genes. However, restricting attention to known pathways will not reveal novel cancer genes or pathways. An alterative strategy is to examine mutated genes in the context of genome-scale interaction networks that include both well characterized pathways and additional gene interactions measured through various approaches. We introduce a computational framework for *de novo* identification of subnetworks in a large gene interaction network that are mutated in a significant number of patients. This framework includes two major features. First, we introduce a diffusion process on the interaction network to define a local neighborhood of "influence" for each mutated gene in the network. Second, we derive a two-stage multiple hypothesis test to bound the false discovery rate (FDR) associated with the identified subnetworks. We test these algorithms on a large human protein-protein interaction network using mutation data from two recent studies: glioblastoma samples from The Cancer Genome Atlas and lung adenocarcinoma samples from the Tumor Sequencing Project. We successfully recover pathways that are known to be important in these cancers, such as the p53 pathway. We also identify additional pathways, such as the Notch signaling pathway, that have been implicated in other cancers but not previously reported as mutated in these samples. Our approach is the first, to our knowledge, to demonstrate a computationally efficient strategy for *de novo* identification of *statistically significant* mutated subnetworks. We anticipate that our approach will find increasing use as cancer genome studies increase in size and scope.

⋆ Supported in part by the "Ing. Aldo Gini" Foundation, Padova, Italy. This work was done while the author was visiting the Department of Computer Science of Brown University.

⋆⋆ BJR is supported by a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

1 Introduction

Cancer is a disease that is largely driven by somatic mutations that accumulate during the lifetime of an individual. Decades of experimental work have identified numerous cancer-promoting oncogenes and tumor suppressor genes that are mutated in many types of cancer. Recent cancer genome sequencing studies have dramatically expanded our knowledge about somatic mutations in cancer. For example, large projects like The Cancer Genome Atlas (TCGA) [31], the Tumor Sequencing Project (TSP) [8], and the Cancer Genome Anatomy Project [11] have sequenced hundreds of protein coding genes in hundreds of patients with a variety of cancers. Other efforts have taken a global survey of approximately 20,000 genes in a 1-2 dozen patients [40, 18, 32]. These studies have shown that: tumors harbor on average approximately 80 somatic mutations; two tumors rarely have the same complement of mutations; and thousands of genes are mutated in cancer [40]. This mutational heterogeneity complicates efforts to distinguish functional mutations from sporadic, passenger mutations. While a few cancer genes are mutated at high frequency (e.g. well known cancer genes like TP53 or KRAS), most cancer genes are mutated at much lower frequencies. Thus, the observed frequency of mutation is an inadequate measure of the importance of a gene, particularly with the relatively modest number of samples that are tested in current cancer studies.

It is widely accepted that cancer is a disease of pathways and it is hypothesized that somatic mutations target genes in a relatively small number of regulatory and signaling networks [12, 39]. Thus, the observed mutational heterogeneity is explained by the fact that there are myriad combinations of alterations that cancer cells can employ to perturb the behavior of these key pathways. The unifying themes of cancer are thus not solely revealed by the individual mutated genes, but by the interactions between these genes. Standard practice in cancer sequencing studies is to assess whether genes that are mutated at sufficiently high frequency significantly overlap known cancer pathways [31, 8, 36, 40, 32, 25].

Finding significant overlap between mutated genes and genes that are members of known pathways is an important validation of existing knowledge. However, restricting attention to these known pathways does not allow one to detect novel group of genes that are members of less characterized pathways. Moreover, it is well known that there is crosstalk between different pathways [39] and dividing genes into discrete pathway groupings limits the ability to detect whether this crosstalk is itself a target of mutations. An additional source of information about gene and protein interactions is large-scale interaction networks, such as the Human Protein Reference Database (HPRD) [22], STRING [17], and others [2, 34]. These resources incorporate both well-annotated pathways and interactions derived from high-throughput experiments, automated literature mining, cross-species comparisons, and other computational predictions. Many researchers have used these interaction networks to analyze gene expression data. Ideker et al. [16] introduced a method to discover subnetworks of differentially expressed genes, and this idea was later extended in different directions by others [30, 26, 38, 21, 28, 13, 5].

We propose to identify "significantly mutated subnetworks" – that is connected subnetworks whose genes have more mutations than expected by chance – *de novo* in a large gene interaction network. This problem differs from the gene expression problem in that a relatively small number of genes might be measured, a small subset of genes in a pathway may be mutated, and that a single mutated gene may be sufficient to perturb a pathway. The naïve approach to *de novo* identification of mutated subnetworks is to examine mutations on all subnetworks, or all subnetworks of a fixed size. This approach is problematic. First, the enumeration of all such subnetworks is prohibitive for subnetworks of a reasonable size. Second, the extremely large number of hypotheses that are tested makes it difficult to achieve statistical significance. Finally, biological interaction networks typically have small diameter due to the presence of "hub" genes of high degree. There are reports that cancer-associated genes have more interaction partners than non-cancer genes [25, 19], and indeed highly mutated cancer genes like TP53 have high degree in most interaction networks (e.g. the degree of TP53 in HPRD is 238). Such correlations might lead to a large number of "uninteresting" subnetworks being deemed significant.

We propose a rigorous framework for *de novo* identification of significantly mutated subnetworks and employ two strategies to overcome the difficulties described above. First, we formulate an *influence* measure between pairs of genes in the network using a diffusion process defined on the graph. This quantity considers a gene to influence another gene if they are both close in distance on the graph *and* there are relatively few paths between them in the interaction network. We use this measure to build a smaller *influence graph* that includes only the mutated genes but encodes the neighborhood information from the larger network. We then identify significant subnetworks using two techniques. The first one requires to solve an NP-hard problem, while in the second one, in which the influence between pairs of genes is enhanced by the number of mutations observed on these genes, the computational problem is reduced to just finding connected components in the graph. Finally, we derive a *two-stage multiple hypothesis test* that mitigates the testing of a large number of hypotheses by focusing on the number of discovered subnetworks of a given size rather than on individual subnetworks. We also show how to estimate the false discovery rate (FDR) associated with this test.

We tested our approach on the HPRD human interaction network using somatic mutation data from two recently published studies: (i) 601 genes in 91 glioblastoma multiforme patients from The Cancer Genome Atlas (TCGA) project; (ii) 623 genes in 188 lung adenocarcinoma patients sequenced during the Tumor Sequencing Project (TSP). In both datasets, we identify statistically significant mutated subnetworks that are enriched for genes on pathways known to be important in these cancers. Our approach is the first, to our knowledge, to demonstrate a computationally efficient strategy for *de novo* identification of *statistically significant* mutated subnetworks. We anticipate that our approach will find increasing use as cancer genome studies increase in size and scope.

2 Methods

In this section we introduce our approach for the identification of significantly mutated pathways in cancer. Due to space constraints, the proofs of theorems are omitted. Supplementary material including details of proofs is available at `http://www.cs.brown.edu/people/braphael/supplements/`.

2.1 Mathematical Model

We model the interaction network by a graph $G = (V, E)$, where the vertices in V represent individual proteins (and their associated genes), and the edges in E represent (pairwise) protein-protein or protein-DNA interactions. Let $\mathcal{T} \subseteq V$ be the subset of genes that have been been tested, or assayed, for mutations in a set $\mathcal{S}$ of samples (patients). The size of $\mathcal{T}$ will vary by study; e.g. some recent works resequenced hundreds of genes [31, 8] while others examine nearly all known protein-coding genes in the human genome [40, 18, 32]. We assume that each gene g is assigned one of two labels, *mutated* or *normal*, in each sample. Let M_i denote the subset of genes in $\mathcal{T}$ that are mutated in the ith sample, for $i = 1, \ldots |\mathcal{S}|$. Let $\mathcal{S}_j$ be the samples in which gene $g_j \in \mathcal{T}$ is mutated, for $j = 1, \ldots, |\mathcal{T}|$, let $m = \sum_i |M_i|$ be the total number of occurrences of altered genes observed in all samples.

We define a *pathway* or *subnetwork* to be a connected subgraph of G. Note that this definition matches the common biological usage of the term where pathways may have arbitrary topology in the graph, and are not restricted to be linear chains of vertices. We generally do not know whether more than one gene must be mutated to perturb a pathway in a sample, and thus will assume that a pathway is mutated in a sample if *any* of the genes in the pathway are mutated. For a subset $T \subseteq \mathcal{T}$, let $S(T)$ denote the set of samples in which *at least one* gene in T is mutated.

2.2 Influence Graph

Our goal is to identify subnetworks that are significant with respect to the set of mutated genes in the samples. The significance of a subnetwork is derived from: (i) the number of samples that have mutations in the genes of the subnetwork, and (ii) the interactions between genes in the subnetwork in the context of the whole network topology. For example, consider two possible scenarios of mutated nodes (Figure 1). In the first scenario, the two mutated nodes are part of a linear chain in the interaction network. In the second scenario, the two mutated nodes are connected through a high-degree node. In the first scenario, there is a single path joining the two mutated nodes and thus we are more surprised by this local clustering of mutations than in the second scenario, where the two nodes are connected by a node that is present in a large number of possible paths

Hubs present an extreme case of this phenomenon and result in many "uninteresting" subnetworks being deemed significant. Since many highly mutated

cancer genes, like TP53, also have high degree in interaction networks it is not advisable to ignore these genes in the analysis of cancer mutation data. These examples show that significance of a subnetwork is derived from both: 1. the number of samples that have mutations in the genes of the subnetwork, and 2. the interactions between genes in the subnetwork in the context of the whole network. A straightforward graph distance like the shortest path between nodes is not sufficient to overcome the problems highlighted above. Moreover, other graph mining approaches like dense subgraph identification [10] are also not appropriate, since not all subnetworks of interest (e.g. the chain in Figure 1) are dense in edges.

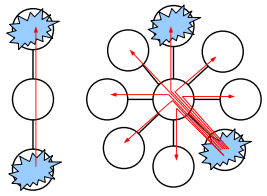

Fig. 1: Mutation on chain vs. star graph.

We use a diffusion process on the interaction network to define a rigorous measure of *influence* between all pairs of nodes. To measure the influence of node s on all the other nodes in the graph, consider the following process, described by [33]. Fluid is pumped into the source node s at a constant rate, and fluid diffuses through the graph along the edges. Fluid is lost from each node at a constant first-order rate γ. Let $f_v^s(t)$ denote the amount of fluid at node v at time t, and let $\mathbf{f}^s(t) = [f_1^s(t), \ldots, f_n^s(t)]^T$ be the column vector of fluid at all nodes. Let L be the Laplacian matrix of the graph[4], and let $L_\gamma = L + \gamma I$. Then the dynamics of this continuous-time process are governed by the vector equation $\frac{d\mathbf{f}^s(t)}{dt} = -L_\gamma \mathbf{f}^s(t) + \mathbf{b}^s u(t)$, where $\mathbf{b}^s$ is the elementary unit vector with 1 at the s^{th} place and 0 otherwise, and $u(t)$ is the unit step function. As $t \to \infty$, the system reaches the steady state. The equilibrium distribution of fluid density on the graph is $\mathbf{f}^s = L_\gamma^{-1}\mathbf{b}^s$ (see [33]). Note that this diffusion process is related to the diffusion kernel [24], or heat kernel [6], which models the diffusion of heat on a graph, and these diffusion processes are also related to certain random walks on graphs [9, 27]. Diffusion processes and their related flow problems have been used in protein function prediction on interaction networks [37, 29] and to define associations between gene expression and phenotype [28].

We interpret f_i^s as the influence of gene g_s on gene g_i. Computing the diffusion process for all tested genes gives us, for each pair of genes $g_j, g_k \in \mathcal{T}$, the influence $i(g_j, g_k)$ that gene g_j has on gene g_k. Note that in general the influence is not symmetric; i.e. $i(g_j, g_k) \neq i(g_j, g_k)$. We define an *influence graph* $G_I = (\mathcal{T}, E_I)$ with the set of nodes corresponding to the set of tested genes, the weight of an edge (g_j, g_k) is given by $w(g_j, g_k) = \min[i(g_k, g_j), i(g_j, g_k)]$. If n is the number of nodes in the interaction network, then the cost of computing G_I is dominated by the complexity of inverting an $n \times n$ matrix.

[4] $L = -A + D$, where A is the adjacency matrix of the graph and D is a diagonal matrix with $D_{i,i} = degree(v_i)$.

2.3 Discovering Significant Subnetworks: Combinatorial Model

Given an influence measure between genes, the obvious first approach for discovering significant subnetworks is to identify sets of nodes in the influence graph G_I that are (1) connected through edges with high influence measure; and (2) correspond to mutated genes in a significant number of samples. We fix a threshold δ and compute a *reduced influence graph* $G_I(\delta)$ of G_I by removing all edges with $w(g_i, g_j) < \delta$, and all nodes corresponding to genes with no mutations in the sample data. The computational problem is reduced to identifying connected subgraphs of $G_I(\delta)$ such that the corresponding set of genes is altered in a significant number of patients.

The size of the connected subgraphs we discover is controlled by the threshold δ. We choose sufficiently small δ such that in the null hypothesis, in which the mutations are randomly placed in nodes corresponding to tested genes, it is unlikely that our procedure finds connected subgraphs with similar properties. Note that value of δ depends only on the null hypothesis and not on the observed sample data (see Section 2.5 for details of the statistical analysis). Finding the connected subgraph of k genes that is mutated in the largest number of samples requires to solve the following problem, that we define as *connected maximum coverage* problem.

Computational Problem Given a graph G defined on a set of m vertices V, a set of elements I, a family of subsets $\mathcal{P} = \{P_1, \ldots, P_m\}$, with $P_i \in 2^I$ associated to $v_i \in V$, and a value k, find the connected subgraph $\mathcal{C}^* = \{v_{i_1}, \ldots, v_{i_k}\}$ with k nodes in G that maximize $|\cup_{j=1}^{k} P_{i_j}|$. In our case we have $G = G_I(\delta)$, V is the subset of genes in $\mathcal{T}$ mutated in at least one sample, and for each $g_i \in V$ the associated set is $\mathcal{S}_i$. The connected maximum coverage problem is related to the maximum coverage problem (see e.g. [14] for a survey) where given a set I of elements, a family of subsets $F \subset 2^I$, and a value k, one needs to find a collection of k sets in F that covers the maximum number of elements in I. This problem is NP-hard as set cover is reducible to it.

If the graph G is a complete graph, the connected maximum coverage problem is the same as the maximum coverage problem. Thus the connected maximum coverage problem is NP-hard for a general graph. Moreover we prove that the problem is still hard even on simple graphs such as the star graph ([35] gives a similar result for the connected set cover problem).

Theorem 1. *The connected maximum coverage problem on star graphs is NP-hard.*

Since the connected maximum coverage problem is NP-hard even for simple graphs we turn to approximate solutions. It is not hard to construct a polynomial time $1 - \frac{1}{e}$ approximation algorithm for spider graphs (analogous to the result in [35] for the connected set cover problem). Since it cannot be applied to the network here, we construct an alternative polynomial time algorithm that gives $O(1/r)$ approximation when the radius of the optimal solution $\mathcal{C}^*$ is r.

Our algorithm obtains a solution $\mathcal{C}_v$ (thus, a connected subgraph) starting from each node $v \in V$, and then returns the best solution found. To obtain

$\mathcal{C}_v$, our algorithm executes an *exploration phase*, i.e. for each node $u \in G$ it finds a shortest path $p_v(u)$ from v to u. Let $\ell_v(u)$ be the set of nodes in $p_v(u)$, and $P_v(u)$ the elements of I that they cover. After this *exploration phase*, the algorithm builds a connected subgraph $\mathcal{C}_v$ starting from v. At the beginning we have $\mathcal{C}_v = \{v\}$. $P_{\mathcal{C}_v}$ is the set of elements covered by the current connected subgraph $\mathcal{C}_v$. Then, while $|\mathcal{C}_v| < k$, the algorithm chooses the node $u \notin \mathcal{C}_v$ such that: $u = \arg\max_{u \in V} \left\{ \frac{|P_v(u) \setminus P_{\mathcal{C}_v}|}{|\ell_v(u) \setminus \mathcal{C}_v|} \right\}$ and $|\ell_v(u) \cup \mathcal{C}_v| \leq K$; the new solution is then $\ell_v(u) \cup \mathcal{C}_v$. The main computational cost of out algorithm is due to the exploration phase, that can be performed in polynomial time. We have the following:

Theorem 2. *The algorithm above gives a* $\frac{1}{cr\frac{2e-1}{e-1}}$*-approximation for the connected maximum coverage problem on* G*, where* $c = \frac{2e-1}{e-1}$ *and* r *is the radius of optimal solution in* G*.*

For our experiments we implemented a variation of this algorithm, that for each pair of nodes (u, v) considers all the shortest paths between u and v, and then keeps the one that maximizes $\frac{|P_v(u)|}{|\ell_v(u)|}$ to build the solution $\mathcal{C}_v$. With this modification the algorithm is not guaranteed to run in polynomial time in the worst-case, but ran efficiently for all our experiments.

2.4 Discovering Significant Subnetworks: the Enhanced Influence Model

We developed an alternative, computationally efficient, approach for identifying subnetworks that are significant with respect to the gene mutation data. The *Enhanced Influence Model* is based on the idea of enhancing the influence measure between genes by the number of mutations observed in each of these genes, and then decomposing an associated *enhanced influence graph* into connected components.

We define the *enhanced influence* graph H. It has a node for each gene g_j with at least one mutation in the data. The weight of edge (g_j, g_k) in H is given by $h(g_j, g_k) = \min\{i(g_j, g_k), i(g_k, g_j)\} \times \max\{|\mathcal{S}_j|, ||\mathcal{S}_k|\}$. Thus, the strength of connection between two nodes in the enhanced influence graph is a function of both the interaction between the nodes in the interaction network and the number of mutations observed in their corresponding genes. Next we remove all edges with weight smaller than a threshold δ to obtain a graph $H(\delta)$. We return the connected components in $H(\delta)$ as the significant subnetworks with respect to the mutation data and the threshold δ. The computational cost is the complexity of computing all connected components in a graph with $|S|$ nodes (number of mutated genes), which is linear in the size of the graph. The significance of the discovered subnetworks depends on the choice of δ. We choose sufficiently small δ such that in the null hypothesis, in which the mutations are randomly placed in nodes corresponding to tested genes, it is unlikely that our procedure finds connected components of similar size (see Section 2.5 for details).

2.5 Statistical Analysis

We assess the statistical significance of our discoveries with respect to null hypothesis distributions in which the mutated genes are randomly allocated in the network, i.e. when the occurrence of mutations are independent of the network topology. We consider two null hypothesis distributions: in H_0^{sample} a total of $m = \sum_i |M_i|$ mutations are placed uniformly at random in the nodes corresponding to the $|\mathcal{T}|$ tested genes. While easier to analyze, this model does not account for the fact that in the observed data a large number of mutations are concentrated in a few genes(e.g. TP53). Thus, we also use a second null hypothesis distribution, H_0^{gene}, generated by permuting the identities of the tested genes in the network. That is we select a random permutation σ of the set $\{1, \ldots, |\mathcal{T}|\}$, and we assign gene g_j, that was mutated in the set of samples $\mathcal{S}_j \subseteq \mathcal{S}$, to the location of gene $g_{\sigma(j)}$ in the original network.

A Two Stage Multi-Hypothesis Test A major difficulty in assessing the statistical significance of the discovered subnetworks is that we test simultaneously for a large number of hypotheses; each connected subnetwork in the interaction graph with at least one tested gene is a possible significant subnetwork and thus an hypothesis. The strict measure of significance level in multi-hypothesis testing is the *Family Wise Error Rate (FWER)*, the probability of incurring at least one Type I error in any of the individual tests. An alternative, less conservative approach to control errors in multiple tests is the the *False Discovery Rate (FDR)* [3]. Let V be the number of Type I errors in the individual tests, and let R be the total number of null hypotheses rejected by the multiple test. We define $FDR = E[V/R]$ to be the expected ratio of erroneous rejections among all rejections (with $V/R = 0$ when $R = 0$). Let h be the total number of hypothesis tested. Applying either measure to our problem, a discovery would be flagged as statistically significant only if its p-value is $O(1/h)$, which is impractical in the size of our problem. Instead, building on an idea presented in [23], we develop a two stage test for our problem that allows us to flag a number of subnetworks in our data as statistically significant with small false discovery rate (FDR) values.

We demonstrate our method through the analysis of the Enhanced Influence model. A similar technique was applied to the Combinatorial model. Let $C_1, \ldots, C_\ell$ be the set of connected components found in the enhanced influence graph $H(\delta)$. Testing for the significance of these discoveries is equivalent to simultaneously testing for $2^{|\mathcal{T}|}$ hypothesis. To reduce the number of hypothesis we focus on an alternative statistic: the *number* of discoveries of a given size. Let $\tilde{r}_s$ be the number of connected components of size $\geq s$ found in the graph $H(\delta)$, and let r_s be the corresponding random variable in the null hypothesis (H_0^{sample} or H_0^{gene}). We are testing now for just $\mathcal{K} = |\mathcal{T}|$ simple hypotheses, for $s = 1, \ldots, \mathcal{K}$: $E_s \equiv$ "$\tilde{r}_s$ conforms with the distribution of r_s". Testing each hypothesis with confidence level $\alpha/\mathcal{K}$, the first stage of our test identifies the smallest size s such that with confidence level α we can reject the null hypothesis that $\tilde{r}_s$ conforms with the distribution of r_s.

The fact that the number of connected components of size at least s is statistically significant does not imply necessarily that each of the connected components is significant. We now add a second condition to the test that guarantees an upper bound on the False Discovery Rate (FDR):

Theorem 3. *Fix $\beta_1, \beta_2, \ldots, \beta_{\mathcal{K}}$ such that $\sum_{i=1}^{\mathcal{K}} \beta_i = \beta$. Let s^* be the first s such that $\tilde{r}_s \geq \frac{\mathbf{E}[r_s]}{\beta_s}$. If we return as significant all connected components of size $\geq s^*$, then the FDR of the test is bounded by β.*

In our tests we have used $\beta_i = \frac{\beta}{2^i}$ for the i^{th} largest s tested (with $\beta_s = \beta - \sum_i \beta_i$ for the smallest s), since we are more interested in finding large connected components.

Estimating the Distribution of the Null Hypothesis. The null hypothesis distributions can be estimated by either a Monte-Carlo simulation ("permutation test") or through analytical bounds.

Using Monte-Carlo simulation, two features of our method significantly reduce the cost of the estimates. First, the Influence Graph G_I is created *without* observing the sample data. The mutation data and G_I are then combined to create the sample dependent graphs $G_I(\delta)$ and $H(\delta)$. Thus, the Monte Carlo simulation needs to run on the graph G_I which is significantly smaller than the original interaction network (in our data the original interaction network had 18796 nodes while the influence graph had only about 600 nodes). Second, our statistical test does not use the p-values of individual connected subgraphs/components but the p-value of the distribution of the number of connected subgraphs/components of a given size. Thus, for this test it is sufficient to estimate p-values that are a magnitude larger, and therefore require significantly fewer rounds of simulations. These features allowed us to compute the null distributions through Monte-Carlo simulations for the size of our data with no significant computational cost. For larger number of tested genes we can estimate the null hypothesis through analytical bounds.

3 Experimental Results

We applied our approach to analyze somatic mutation data from two recent studies. The first dataset is a collection of 453 somatic mutations identified in 601 tested genes from 91 glioblastoma multiforme (GBM) samples from The Cancer Genome Atlas [31]. In total, 223 genes were reported mutated in at least one sample. The second dataset is a collection of 1013 somatic mutations identified in 623 tested genes from 188 lung adenocarcinoma samples from the Tumor Sequencing Project [8]. In total, 356 genes were reported mutated in at least one sample. For the Enhanced Influence model we also considered simulated data.

We use the protein interaction network from the Human Protein Reference Database (June 2008 version) [22] which consists of 18796 vertices and 37107

edges. We derive the influence graph for each dataset by directly computing the inverse[5] of L_γ. The results presented below are obtained by fixing the parameter $\gamma = 8$, which is approximately the average degree of a node in HPRD (after the removal of disconnected nodes). We also considered $\gamma = 1$ and $\gamma = 30$: in both cases the results obtained are close to the ones obtained with $\gamma = 8$.

The resulting influence graphs have weights $i(g_j, g_k) \neq 0$ for almost all pairs (g_j, g_k) of tested genes: less than 2% of the weights are zero in the GBM graph, while all weights in the lung adenocarcinoma graph are positive. Supplementary tables are available at `http://www.cs.brown.edu/people/braphael/supplements/`.

3.1 Combinatorial Model

We used the combinatorial model to extract a subnetwork of k mutated genes that is mutated in the highest number of samples from GBM and lung adenocarcinoma with $k = 10$ and $k = 20$. For both datasets we used the procedure described in Section 2.3 to derive the threshold $\delta = 0.0001$ for the reduced influence graph $G_I(\delta)$. Table 1 shows that we find statistically significant subnetworks under both the H_0^{gene} and H_0^{sample} null hypotheses (p-values for H_0^{sample} are computed without Monte-Carlo simulation). To assess the biological significance of our findings in GBM, we compared the genes in each subnetwork to the genes in pathways that were previously implicated in GBM and used as a benchmark in the TCGA publication [31] (See also Figure 2 (a) below). We find that our subnetworks are enriched for genes in the RTK/RAS/PI(3)K pathway and to a lesser extent, the p53 pathway. For the lung adenocarcinoma samples, we find that the subnetworks share significant overlap with the pathways reported in the original publication [8]. These results demonstrate that the combinatorial model is effective in recovering genes known to be important in each of these cancers.

3.2 Enhanced Influence Model

Simulated Data. We tested the ability of our enhanced influence model to recover significantly mutated pathways in simulated data. We extracted a well-curated network of 258 genes called "Pathways in cancer (hsa05200)" from the KEGG database [20]. We augmented this network with additional random edges so that 20% of the edges of the resulting network were random. We assigned mutations to a well-known cancer signaling pathway, PKC -RAF - MEK - ERK, a linear chain $\mathcal{P}$ of 4 genes, so that at least one gene is mutated in x% of samples, for different x. We then randomly assigned mutations to all the genes in the network matching the observed values (e.g. number of samples, ratio between number of tested genes and number of genes in the network, etc.) We correctly identify $\mathcal{P}$ as significantly mutated ($P < 10^{-2}$, FDR $< 10^{-2}$) even when each gene in $\mathcal{P}$ is altered in ≤ 5% of the samples, but $\mathcal{P}$ is altered in 17% of the samples.

[5] In contrast [33] derive a power series approximation to L_γ^{-1} whose convergence depends on the choice of γ.

			p-val		pathway enrichment p-val		
dataset	k	samples	H_0^{sample}	H_0^{gene}	all	RTK/RAS/PI(3)K	p53
GBM	10	67	$< 10^{-10}$	4×10^{-3}	3×10^{-4}	8×10^{-4}	0.19
	20	78	$< 10^{-10}$	$< 10^{-3}$	10^{-5}	8×10^{-5}	0.05
Lung	10	140	$< 10^{-10}$	0.02	8×10^{-6}	/	
	20	151	$< 10^{-10}$	0.03	3×10^{-3}	/	

Table 1: Results of the combinatorial model. k is the number of genes in the subnetwork. *samples* is the number of samples in which the subnetwork is mutated. *p-val* is the probability of observing a connected subgraph of size k under the random model H_0^{sample} or H_0^{gene}. *enrichment p-val* is the p-value of the hypergeometric test for overlap between genes in the identified subgraph and genes reported significant pathways in [31] or [8]. For GBM, *enrichment p-val* is the p-value of the hypergeometric test for RTK/RAS/PI(3)K and p53 pathways.

Note that genes mutated in 5% of the samples were not reported as significantly mutated in [31], demonstrating that our method correctly identifies a mutated path even when the individual genes in the path are not mutated in a significant number of samples. Moreover, $\mathcal{P}$ is the *only* significant pathway reported by our method. To verify that our influence measure takes into account the topology of the network, we added a number of edges to the RAF gene in $\mathcal{P}$, giving it high degree in the network. As expected, $\mathcal{P}$ is no longer identified as significant in the modified network.

Real data. We applied the enhanced influence model to the GBM and lung adenocarcinoma datasets. Following the procedure described in Section 2.4, we first computed the enhanced influence network, using a threshold of $\delta = 0.003$ for the GBM data and $\delta = 0.01$ for the lung adenocarcinoma data. Table 2 shows the number and sizes of the connected components identified in the GBM data, and the associated p-values, the latter obtained using the method described in Section 2.5. We identify two significant connected components with more than 19 genes (FDR ≤ 0.14). We find significant overlap ($P < 10^{-2}$ by hypergeometric test) between the 68 genes in our connected components and the set of all mutated genes in the same RTK/RAS/PI(3)K, p53, and RB pathways examined in the TCGA study [31]. The second largest connected component with 19 genes has significant overlap to the p53 pathway, while the largest connected component with 22 genes has significant overlap with the RTK/RAS/PI(3)K signaling pathway. In contrast to the combinatorial model, the enhanced influence model separates these two pathways into different connected components. Figure 2 (a) illustrates the overlap between the mutated genes in connected components returned by our method and genes in the pathways reported in [31].

For the lung data, Table 3 shows the sizes of connected components returned by the enhanced influence model and the p-values associated with each. The 88 genes in the union of the connected components derived by our method overlap significantly ($P < 7 \times 10^{-9}$ by the hypergeometric test) with the mutated

		H_0^{sample}		H_0^{gene}		enrichment p-val	
s	# c.c. $\geq s$	μ	p-val	μ	p-val	RTK/RAS/PI(3)K	p53
2	15	22.18	0.97	13.63	0.38	/	/
3	3	6.37	0.98	4.38	0.6	/	/
19	2	$< 10^{-3}$	$< 10^{-3}$	0.07	$< 10^{-3}$	0.9	4×10^{-3}
22	1	$< 10^{-3}$	$< 10^{-3}$	0.05	0.05	4×10^{-6}	–

Table 2: Results of the enhanced influence model on GBM samples. s is the size of connected components (c.c.) found with our method. # *c.c.* $\geq s$ is the number of c.c. with *at least* s nodes. μ is the expected number of c.c. with $\geq s$ nodes under random models H_0^{gene}, H_0^{sample}. *p-val* is the probability of observing *at least* # *c.c.* $\geq s$ with at least s nodes in a random dataset. The last 3 columns show, for c.c. with $s > 3$, the result of the hypergeometric test for enrichment for RTK/RAS/PI(3)K, and p53 pathways respectively.

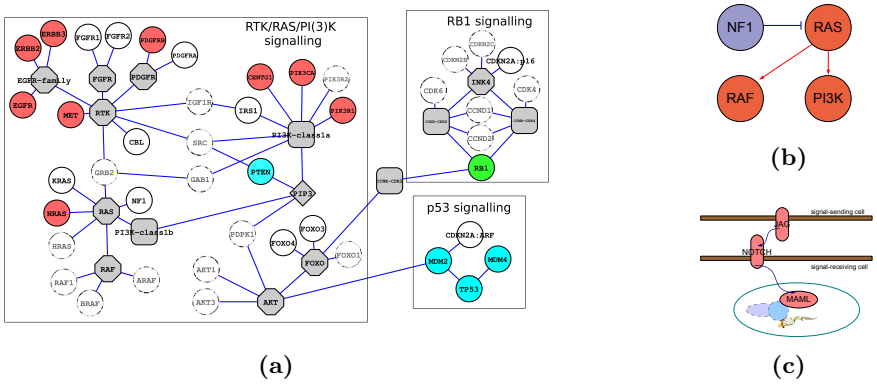

Fig. 2: **(a)** Overlap between subnetworks found by the enhanced influence model and significant pathways reported in [31]. Each circle is a gene, gray nodes represents protein families or complexes, or small molecules. For each protein family and complex, tested genes are shown. "Dashed" nodes are tested genes that were not mutated in GBM, and thus cannot be returned as significant. Red nodes are found in the c.c. of size 22, blue nodes in the c.c. of size 18, and the green node in a c.c. of size 2. **(b)** Pathway corresponding to one of the connected components extracted with enhanced influence model in lung. **(c)** Notch signaling pathway identified in the lung dataset.

		H_0^{sample}		H_0^{gene}		
s	# c.c. $\geq s$	μ	p-val	μ	p-val	enrichment p-val
2	24	23.4	0.7	17.67	0.4	/
3	11	6.51	0.13	7.27	0.2	/
4	7	3.21	0.07	4.98	0.13	/
5	5	2.09	0.01	2.18	0.01	/
7	4	0.54	0.01	0.56	0.01	–
10	3	$< 10^{-3}$	$< 10^{-3}$	0.4	0.02	0.34; 10^{-5}; 9×10^{-8}

Table 3: Results of the enhanced influence model on lung adenocarcinoma samples. Columns are as described in Table 2. Last column shows, for c.c. with $s \geq 7$, the result of the hypergeometric test for enrichment all genes reported in significant pathways in [8] (the 3 values shown refers to c.c. of size 10).

pathways reported in the network of Figure 6 in the TSP publication [8]. We identify 4 connected components of size ≥ 7 (FDR ≤ 0.56). The first connected component of size 10 contains genes in the p53 pathway, and the second one is enriched ($P < 10^{-2}$) for the MAPK pathway (Figure 2 (b)). The third component is the ephrin receptor gene family, a large family of membrane-bound receptor tyrosine kinases, that were reported as mutated in breast and colorectal cancers [36]. Notably, only one of the genes in this component, EPHA3, is mentioned as significantly mutated in [8]. Finally, the connected component of size 7 consists exclusively of members of the Notch signaling pathway (Figure 2 (c)). The mutated genes include: the Notch receptor (NOTCH2/3/4); Jagged (JAG1/2), the ligand of Notch; and Mastermind (MAML1/2), a transcriptional co-activator of Notch target genes. The Notch signaling pathway is a major developmental pathway that has been implicated in a variety of cancers [1] including lung cancer [7]. Mutations in this pathway were not noted in the original TSP publication [8], probably because no single gene in this pathway is mutated in more than 3 samples. Because our method exploits both mutation frequency and network topology, we are able to identify these more subtle mutated pathways, and in this case identify an entire "signaling" circuit.

3.3 Naïve Approach

To demonstrate the impact of the influence graph on the results, we implemented a naïve approach that examines all paths in the original HPRD network that connect two tested genes and contain at most 3 nodes. We extracted all paths that were altered in a significant number of samples with FDR ≤ 0.01 using the standard Benjamini-Yekutieli method [4]. More than 1700 paths in GBM and > 2200 in lung adenocarcinoma are marked as significant with this method. A major reason for this large number of paths is the presence of highly mutated genes that are also high-degree nodes in the HPRD network (e.g. TP53). *Each* path through these high degree nodes is marked as significant. One possible solution is to remove any path that contains a subpath that is significant How-

ever, these filtered paths include *none* through important highly-mutated and high degree genes (like TP53). Our influence graph uses both mutation frequency and local topology of the network, allowing us to recover subnetworks containing these genes. Finally, we note that finding larger, statistically significant subnetworks (e.g. those with 10 or 20 nodes) with the naïve approach is impossible in the GBM and lung datasets because of the severe multiple hypotheses correction for the large number of subnetworks tested; e.g., the number of connected components with 10 tested nodes in the HPRD network is $> 10^{10}$. For the same reason the enumeration of all the paths or connected components of reasonable size is impossible.

4 Discussion

We present an approach to identify significantly mutated pathways in a large, unannotated interaction network. The subnetworks derived by our method share significant overlap with the known cancer pathways such as the manually curated pathways in TCGA [31]. Remarkably, we automatically extracted a large fraction of these pathways with modest number (100-200) of samples (Figure 2). Our approach has two key advantages over the common strategy of testing the overlap between mutated genes and genes from known pathways approach, using a hypergeometric or similar test. First, we incorporate biological information that is not presently represented in existing well-characterized pathways, while accounting for the uncertainty in large gene interaction networks. Second, we are able to assign significance to genes that are altered at low frequency but are part of a larger subnetwork that is altered at significant frequency. The latter advantage was demonstrated in the lung adenocarcinoma dataset where we identify the Notch signaling pathway as significant, even though the individual genes were not mutated at significant frequency.

We plan to extend our model in numerous directions, including: (i) inclusion of other types of mutations such as copy number changes in genes, genome rearrangements, gene expression, or epigenetic alterations; (ii) extension of the interaction network to include additional interaction types (e.g. regulatory or miRNA) as well as directed interactions (activating vs. inhibitory); (iii) consideration of errors in the interaction network. The later can be included naturally in our diffusion model by adding weights, or reliabilities, on the edges. Moreover, we have adapted our model to take into account the length of the genes in the network, weighting the frequency of mutation in a gene by its length. The results obtained for the GBM and lung adenocarcinoma data are extremely close to the one presented here (data not shown).

We anticipate that our method will become even more useful as larger datasets become available. Several recent studies [40, 18, 32] have surveyed a much larger number of genes than considered here (approximately 20,000), but in a relatively small number of samples (1-2 dozen per cancer type). Continuing decline in sequencing costs and the development of targeted exon-capture techniques [15] will soon enable global surveys of all protein-coding genes in hundreds to thousands of cancer samples.

References

1. H. Axelson. Notch signaling and cancer: emerging complexity. *Semin. Cancer Biol.*, 14:317–319, 2004.
2. G. D. Bader, I. Donaldson, C. Wolting, B. F. Ouellette, T. Pawson, and C. W. Hogue. BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res.*, 29:242–245, Jan 2001.
3. Y. Benjamini and Y. Hochberg. Controlling the false discovery rate. *J. Royal Statistical Society*, Series B(57):289–300, 1995.
4. Y. Benjamini and D. Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of Statistics*, 29(4):1165–1188, 2001.
5. H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, 3:140, 2007.
6. F. Chung. The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, 104(50):19735, 2007.
7. B. J. Collins, W. Kleeberger, and D. W. Ball. Notch in lung development and lung cancer. *Semin. Cancer Biol.*, 14:357–364, 2004.
8. L. Ding et al. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–75, 2008.
9. P.G. Doyle and J.L. Snell. *Random Walks and Electric Networks*. The Mathematical Association of America, 1984.
10. U. Feige, G. Kortsarz, and D. Peleg. The dense k-subgraph problem. *Algorithmica*, 29:2001, 1999.
11. C. Greenman et al. Patterns of somatic mutation in human cancer genomes. *Nature*, 446:153–158, 2007.
12. W. C. Hahn and R. A. Weinberg. Modelling the molecular circuitry of cancer. *Nat Rev Cancer*, 2(5):331–41, 2002.
13. B. J. Hescott, M. D. M. Leiserson, L. Cowen, and D. K. Slonim. Evaluating between-pathway models with expression data. In *RECOMB*, pages 372–385, 2009.
14. D. S. Hochbaum, editor. *Approximation algorithms for NP-hard problems*. PWS Publishing Co., Boston, MA, USA, 1997.
15. E. Hodges et al. Genome-wide in situ exon capture for selective resequencing. *Nat. Genet.*, 39:1522–1527, 2007.
16. T. Ideker, O. Ozier, B. Schwikowski, and A. F. Siegel. Discovering regulatory and signalling circuits in molecular interaction networks. *Bioinformatics*, 18 Suppl 1:S233–240.
17. L. J. Jensen et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.*, 37:D412–416, 2009.
18. S. Jones et al. Core signaling pathways in human pancreatic cancers revealed by global genomic analyses. *Science*, 321(5897):1801–6, 2008.
19. P. F. Jonsson and P. A. Bates. Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22:2291–2297, 2006.
20. M. Kanehisa and S. Goto. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.
21. S. Karni, H. Soreq, and R. Sharan. A network-based method for predicting disease-causing genes. *J. Comput. Biol.*, 16:181–189, 2009.
22. T. S. Keshava Prasad et al. Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, 37:D767–772, 2009.
23. A. Kirsch, M. Mitzenmacher, A. Pietracaprina, G. Pucci, E. Upfal, and F. Vandin. An efficient rigorous approach for identifying statistically significant frequent itemsets. In *PODS*, pages 117–126, 2009.

24. Risi Imre Kondor and John Lafferty. Diffusion kernels on graphs and other discrete structures. In *In Proceedings of the ICML*, pages 315–322, 2002.
25. J. Lin et al. A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Res.*, 17:1304–1318, 2007.
26. M. Liu et al. Network-based analysis of affected biological processes in type 2 diabetes models. *PLoS Genet.*, 3:e96, 2007.
27. L. Lovász. Random walks on graphs: A survey, 1993.
28. X. Ma, H. Lee, L. Wang, and F. Sun. CGI: a new approach for prioritizing genes by combining gene expression and protein-protein interaction data. *Bioinformatics*, 23:215–221, 2007.
29. E. Nabieva, K. Jim, A. Agarwal, B. Chazelle, and M. Singh. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1:i302–310, 2005.
30. S. Nacu, R. Critchley-Thorne, P. Lee, and S. Holmes. Gene expression network analysis and applications to immunology. *Bioinformatics*, 23:850–858, 2007.
31. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–8, 2008.
32. D. W. Parsons et al. An integrated genomic analysis of human glioblastoma multiforme. *Science*, 321(5897):1807–12, 2008.
33. Y. Qi, Y. Suhail, Y. Y. Lin, J. D. Boeke, and J. S. Bader. Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res.*, 18:1991–2004, 2008.
34. L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg. The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, 32:D449–451, Jan 2004.
35. T.-P. Shuai and X.-D. Hu. Connected set cover problem and its applications. In *AAIM*, pages 243–254, 2006.
36. T. Sjoblom et al. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314(5797):268–74, 2006.
37. K. Tsuda and W. S. Noble. Learning kernels from biological networks by maximizing entropy. *Bioinformatics*, 20 Suppl 1:i326–333, 2004.
38. I. Ulitsky, R. M. Karp, and R. Shamir. Detecting disease-specific dysregulated pathways via analysis of clinical expression profiles. In *RECOMB*, pages 347–359, 2008.
39. B. Vogelstein and K. W. Kinzler. Cancer genes and the pathways they control. *Nat. Med.*, 10:789–799, 2004.
40. L. D. Wood et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–13, 2007.