Performance Metrics for Group-Detection Algorithms

J.V. White, S. Steingold, C.G. Fournelle
Intelligent Systems Division
ALPHATECH, Inc.
Burlington, MA 01803

Abstract

A group-detection algorithm attempts to identify groups of entities in relational data that belong to specific groups or subsystems, based on records of interactions among small subsets of the entities. For example, such algorithms may be used to detect groups (or systems) of interacting proteins in bio-networks based on multiple experiments, where each experiment attempts to identify only a small subset of the studied system. Measurements are typically noisy because they contain extraneous entities that are not members of the groups being studied. Therefore, a statistical characterization of group-finding performance is needed. This paper discusses metrics for measuring the probabilistic performance of group-detection algorithms. The metrics may be used to compare algorithms and to assess their performance in Monte Carlo simulation studies. We show that several traditional performance metrics are deficient if the size of a group is very small compared to the size of the population of entities being considered. Moreover, a pair of classical metrics (such as sensitivity and specificity or recall and precision) must be used to track the two types of errors. To address these two issues, a new information-theoretic metric, termed *proficiency*, is introduced. Proficiency may be used to measure the performance of any detection algorithm, including classical hypothesis tests in statistics.

1 Introduction

Databases containing information relating entities to each other, whether those entities are proteins, gene products, chemicals, or quite different entities such as grocery store items or recipe ingredients, are ubiquitous. Because of the large volumes of data available today for analysis and research, automated techniques for processing vast amounts of information are gaining increasing attention. Moreover, gains in the fields of automated information extraction and data analysis enable these datasets to grow at an incredible speed, drowning a data analyst in a sea of information. Automated link-analysis systems have emerged that ingest the relational content of these databases and attempt to provide insight into underlying relationships between the entities, for example [3, 4, 5, 7, 8]. This paper addresses one method of link analysis called *group detection*, which is directly applicable to the study of molecular interactions in gene, protein, and metabolite networks. Group-detection algorithms attempt to identify subsets of entities that exhibit a significant level of interconnectivity or interaction based on multiple pieces of noisy fragmentary experimental evidence, with each piece linking a subset of the entities together. Depending on the domain of the data set, the groups may suggest systems of interacting proteins in a bio-network or items likely to be purchased together by a consumer.

Group detection differs from traditional clustering because each entity may belong to any

number of groups. However, group detection may be viewed as a collection of detection problems or hypothesis tests, with one test or detector for each group. Thus, the performance of a group-detection algorithm may be quantified using the classical metrics of detection theory. Unfortunately, many of the traditional detection-system metrics become insensitive to one or the other of the two types of errors (false positives and false negatives) if the group sizes are very small compared to the number of entities being considered. Moreover, two classical metrics must be considered simultaneously (such as sensitivity and specificity or recall and precision). To address these shortcomings, we introduce a new metric derived from information theory, which we term Proficiency. Proficiency is a scalar metric, based on mutual information and entropy, that characterizes a detection system's performance and avoids some of the limitations of the traditional performance metrics.

Section 2 provides a mathematical formulation of the the group-detection problem. Section 3 discusses traditional performance metrics and introduces the proficiency metric. Section 4 provides an illustration of how we have used the proficiency metric to study the performance of a particular group-detection algorithm, K-GROUPS [6], using simulated data, and Section 5 provides a summary. The mathematical symbols are listed in Appendix A.

2 Mathematical formulation

Consider a population of N_E entities. Each entity may belong to any of N_G groups. The members of group k interact with each other because of their common group membership. Those entities who belong to none of the groups are called *orphans*. A vector $\mathbf{x}_g$ of indicator variables is defined for each group: $\mathbf{x}_g(e) = 1$ if the entity e is in the group g; and $\mathbf{x}_g(e) = 0$ otherwise. We model the groups as being exchangeable and the entities as being exchangeable. That is, any joint probability distribution on the groups is invariant under a reordering of the groups and the same for entities. Therefore, the probability that any particular entity belongs to any particular group is some number, say P_G. To model group membership, we use the random variable $\mathcal{X}(P_G)$, which takes the value 1 with probability P_G and the value 0 with probability $1 - P_G$.

We now introduce a generative model for observations of interactions between entities, which is a special case[1] of the generative model on which the GDA detection algorithm [6] is based. A link vector $\mathbf{L} = \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix}$ is a list of n distinct entities observed to interact with each other on some occasion. We model each link vector as a sample $\mathbf{L} \sim \mathcal{L}$ from a parametric family of random variables $\{\mathcal{L}(n, N_G, P_b, P_r)\}$, defined as follows. Parameter n is the number of entities in the link; N_G is the number of groups; P_b is the probability that the link is generated by an interaction *between members* of some group, where all N_G groups are equally probable, and P_r is the probability that any entity in a group-generated link is actually not a group member. With probability $1 - P_b$ the link is generated as a random sample from the entire population without replacement. If the link is generated by an interaction within a group, then each entity in the link is selected as follows: with probability P_r, the entity is a random selection from outside the group; and with probability $1 - P_r$, the entity is a random selection from within the group. All sampling within a link vector occurs without replacement so that all entities in the link are distinct.

A group-detection algorithm $A(q)$ takes an integer q and a set of link vectors as inputs and then outputs q decision vectors $\{\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_q\}$ with $\mathbf{y}_g \in \{0, 1\}^{N_E}$. The decision

[1] The link vectors, in this report, all have the same size and type and all entities belong to a single demographic class.

vector $\mathbf{y}_g$ specifies which entities the algorithm has assigned[2] to its gth group: $\mathbf{y}_g(e) = 1$ if the entity e is assigned to group g; and $\mathbf{y}_g(e) = 0$ otherwise.

This setup may be depicted as

$$\overbrace{\mathcal{X}(P_G) \xrightarrow{\text{Sample}} \underbrace{\begin{bmatrix}\mathbf{x}_1\\ \mathbf{x}_2\\ \vdots\\ \mathbf{x}_{N_G}\end{bmatrix} \longrightarrow \mathcal{L}(n, N_G, P_b, P_r) \xrightarrow{\text{Sample}} \begin{bmatrix}\mathbf{L}_1\\ \mathbf{L}_2\\ \vdots\\ \mathbf{L}_m\end{bmatrix}}_{\text{Link Generation}} \longrightarrow A(q) \longrightarrow \begin{bmatrix}\mathbf{y}_1\\ \mathbf{y}_2\\ \vdots\\ \mathbf{y}_q\end{bmatrix}}^{\text{Group Generation}} \qquad \overbrace{\qquad\qquad}^{\text{Detection System}}.$$

Since we are discussing exchangeable groups and exchangeable entities, the probabilistic performance of the detector is determined by the 2×2 joint probability matrix of x and y, which is defined as $\mathbf{P}_{xy}(j, k) = \Pr[\mathbf{x}_g(e) = j, \mathbf{y}_g(e) = k]$ for $e \in \{1, 2, \ldots, N_E\}$ and $g \in \{1, 2, \ldots, N_G\}$. From now on, we use x and y to denote $\mathbf{x}_g(e)$ and $\mathbf{y}_g(e)$ for any e and g.

3 Probabilistic performance metrics

A probabilistic performance metric is a function that maps $\mathbf{P}_{xy}$ into a real number. These metrics typically measure error rates, success rates, information rates, expected utility, or expect cost. Since the elements of $\mathbf{P}_{xy}$ sum to 1, every performance metric may be defined as a function of three parameters. Any three parameters that uniquely determine $\mathbf{P}_{xy}$ may be used. For example, the following parameters are widely used.

Prevalence of group members The group prevalence $P_G = \Pr[x = 1]$, also known as the *prior probability of group membership*. This parameter measures the expected fraction of the population belonging to a particular group.

True-positive rate The true-positive rate $T_p = \Pr[y = 1|x = 1]$, also known as the *detection probability*, *sensitivity*, and *recall metric*. This parameter is the probability that the detector will assign an entity to a particular group, given that the entity is a member of that group. The complement of this metric is the *false-negative rate* $F_n = 1 - T_p$, which is also known as the *miss rate*.

False-positive rate The false-positive rate $F_p = \Pr[y = 1|x = 0]$, also known as the false-alarm probability. This is the probability that a detector will assign an entity to a particular group, given that the entity is not in that group. The complement of this metric is the *true-negative rate* $T_n = 1 - F_p$.

For convenient reference, our notation is summarized in Table 3. The joint probability matrix satisfies

$$\mathbf{P}_{xy} = \begin{bmatrix}(1 - P_G)T_n & (1 - P_G)F_p\\ P_G F_n & P_G T_p\end{bmatrix},$$

[2]For Monte Carlo simulation studies based on simulated link vectors, the groups specified by A may be assumed to be numbered from 1 to q so as to optimize a measure of similarity between these groups and the actual groups in the population. That is, $\{(\mathbf{y}_1, \mathbf{x}_1), (\mathbf{y}_2, \mathbf{x}_2), \ldots\}$ is an optimal assignment of $\mathbf{y}$'s to $\mathbf{x}$'s. In simulation studies, we use the proficiency metric as the measure of similarity and optimize the group assignments by using the Hungarian algorithm [1].

and the marginal probabilities are

$$\mathbf{P}_x = \begin{bmatrix} (1-P_G) \\ P_G \end{bmatrix},$$
$$\mathbf{P}_y = \begin{bmatrix} (1-P_G)T_n + P_G F_n \\ (1-P_G)F_p + P_G T_p \end{bmatrix}.$$

Moreover, the parameters P_G, T_p, and F_p may be expressed as functions of $\mathbf{P}_{xy}$ and $\mathbf{P}_x$,

$$P_G = \mathbf{P}_x(2),$$
$$T_p = \mathbf{P}_{xy}(2,2)/\mathbf{P}_x(2),$$
$$F_p = \mathbf{P}_{xy}(1,2)/\mathbf{P}_x(1).$$

Comment The following subsections discuss well known classical performance metrics. A shortcoming of these metrics is that they all become insensitive to either F_n or F_p as the probability of group membership P_G goes to 0. For assessing performance in Monte Carlo simulation studies, this shortcoming leads to complications because, to track performance, we must consider *two* of these metrics simultaneously, such as recall and precision. What is missing among the classical metrics is a *single* metric that (1) has a nice interpretation and (2) retains sensitivity to both F_n and F_p in the limit $P_G \to 0$. Such a performance metric, termed *proficiency*, will be defined based on information-theoretic measures.

3.1 Error rate

The *error rate* P_E is the probability that the detector will incorrectly classify an entity: $P_E = \Pr[x = 1, y = 0] + \Pr[x = 0, y = 1]$, which equals $P_G F_n + (1 - P_G)F_p$. If $0 < P_G < 1$, *ideal performance* is achieved (the error rate is zero) if, and only if, the false-negative rate F_n and the false-positive rate F_p are both zero. The Appendix contains plots of the error rate as a function of F_n and F_p for selected values of the group prevalence P_G.

As group membership becomes very improbable ($P_G \to 0$), the error rate $P_E \to F_p$. Therefore, P_E has the disadvantage of being blind to the miss rate F_n if $P_G << 1$.

3.2 Two predictive values

The *positive predictive value* PV^+ is defined as the conditional (posterior) probability that x is 1, given that y is 1; it satisfies $PV^+ = T_p P_G/\mathbf{P}_y(2)$ and is also known as the *precision* metric. If $P_G \to 0$, then $PV^+ \to (1 - F_n)P_G/F_p$. Therefore, the precision becomes insensitive to small miss rates $F_n << 1$ if $P_G << 1$.

The *negative predictive value* PV^- is defined as the conditional (posterior) probability that x is 0, given that y is 0, and it satisfies $PV^- = (1 - F_p)(1 - P_G)/\mathbf{P}_y(1)$. If $P_G \to 0$, then $PV^- \to (1 - F_p)/F_p$. Therefore, PV^- also becomes insensitive to the miss rate F_n if $P_G << 1$.

3.3 Bayes factors and signal-to-noise ratios

The Bayes factor G_1 favoring $x = 1$, given $y = 1$, is defined by its appearance in the odds-form of Bayes's rule, $O_1^+ = G_1 O_1^-$. Here $O_1^- = P_G/(1 - P_G)$ is the *prior odds*[3] *favoring* $x = 1$ (prior to knowing the value of y), and $O_1^+ = PV^+/(1 - PV^+)$ is the *posterior*

[3]It is useful to note that the odds favoring an event are nearly equal to the probability of the event if this probability is small compared to 1.

odds favoring $x = 1$, given $y = 1$. So G_1 is the factor of proportionality between the prior and the posterior odds O_1^- and O_1^+. This metric satisfies $G_1 = T_p/F_p \equiv (1 - F_n)/F_p$, which is independent of the group prevalence P_G. However, if $F_n \to 0$, then $G_1 \to 1/F_p$. Therefore, G_1 is insensitive to $F_n << 1$.

Another way of looking at G_1 is to consider the group signal-to-noise ratios (SNR_{in} and SNR_{out}) at the input and output of the detector. The input group SNR_{in} is defined as the prior expected number of true positives divided by the prior expected number of false positives (per group). The output group SNR_{out} is defined in like manner with the expectations conditioned on observing $y = 1$ at the detector output. Therefore, the following relations hold,

$$\text{SNR}_{\text{in}} = \frac{N_E P_G}{N_E(1 - P_G)} \equiv O_1^-, \tag{1}$$

$$\text{SNR}_{\text{out}} = \frac{T_p N_E P_G}{F_p N_E(1 - P_G)} \equiv G_1 \text{SNR}_{\text{in}} \equiv G_1 O_1^- \equiv O_1^+. \tag{2}$$

This exposes three useful facts: (1) The prior odds favoring group membership equals SNR_{in}; (2) the posterior odds favoring group membership equals SNR_{out}; and (3) the Bayes factor G_1 measures the change in signal-to-noise ratio provided by the detector.

The other Bayes factor, G_0, favors $x = 0$, given $y = 0$. It is also defined by its appearance in the odds-form of Bayes's rule, $O_0^+ = G_0 O_0^-$, this time favoring $x = 0$, given $y = 0$. Here $O_0^- = (1 - P_G)/P_G$ is the *prior odds favoring* $x = 0$, and $O_0^+ = PV^-/(1 - PV^-)$ is the *posterior odds favoring* $x = 0$, given $y = 0$, where $PV^- = \Pr[x = 0|y = 0]$ is the negative predictive value. This Bayes factor satisfies $G_0 = (1 - F_p)/(1 - T_p) \equiv (1 - F_p)/F_n$. If $F_p \to 0$, then $G_0 \to 1/F_n$, which focuses attention on the false-negative rate. Therefore, G_0 is insensitive to $F_p << 1$.

3.4 Proficiency

The previously considered performance metrics become insensitive to either F_p or F_n as the group prevalence $P_G \to 0$. To find a performance metric that remains sensitive to both F_p and F_n in the limit as P_G approaches zero, we use standard results from information theory [2].

An ideal detector is defined as one for which $F_p = F_n = 0$. Such a detector would provide H_x bits[4] (per entity per group) of information about x, where $H_x = -P_G \log_2 P_G - (1 - P_G) \log_2(1 - P_G)$ is the *entropy of the prior distribution* $\mathbf{P}_x$ on group membership for a randomly selected entity. However, the detector output actually provides only I bits per entity per group, where

$$I = \sum_{j=1}^{2} \sum_{k=1}^{2} \mathbf{P}_{xy}(j, k) \log_2 \frac{\mathbf{P}_{xy}(j, k)}{\mathbf{P}_x(j)\mathbf{P}_y(k)} \tag{3}$$

is the *mutual information between x and y*. From this definition, it follows that I satisfies

$$I = (1 - P_G)T_n I_{00} + (1 - P_G)F_p I_{01} + P_G F_n I_{10} + P_G T_p I_{11},$$

where the *specific information* I_{jk} for event $(x = j, y = k)$ is defined as

$$I_{jk} = \log_2 \frac{\Pr[x = j, y = k]}{\Pr[x = j]\Pr[y = k]} \equiv \log_2 \frac{\mathbf{P}_{xy}(j + 1, k + 1)}{\mathbf{P}_x(j + 1)\mathbf{P}_y(k + 1)}. \tag{4}$$

[4]The standard unit for measuring information is the *bit*, which corresponds to using base-2 logarithms in the definition of entropy and mutual information. If natural logarithms to the base $e = 2.7182\ldots$ are used, then the unit called a *nat*.

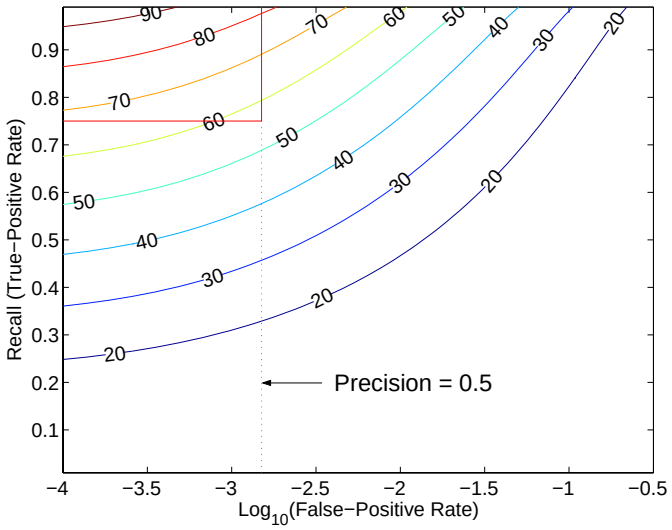

Figure 1: ROC curves for different values of the proficiency metric expressed in percent, $\mathrm{SNR}_{\mathrm{in}} = 0.002$.

Both the entropy and the mutual information are small numbers compared to 1 bit if the probability of group membership is small, $P_G << 1$. For example if P_G is 0.001, then $I \leq H_x = 0.011$. In fact, the smaller the group membership probability P_G (and, therefore, the lower $\mathrm{SNR}_{\mathrm{in}}$ is, and the harder it is to find members of the group), the smaller I becomes. To measure the extraction of useful information about group membership in an intuitive way, we define the *proficiency* metric $\alpha = I/H_x$, which is the fraction of information needed for ideal detection that is actually delivered by the detection algorithm. The Appendix contains contour plots of proficiency as a function of F_n and F_p for selected values of P_G.

The definition of proficiency requires some care if the true group is either empty ($P_G = 0$) or contains all the entities ($P_G = 1$). Such a group is termed *nonrandom*, and its entropy H_x is zero. In contrast, a *random* group is defined as one having $0 < P_G < 1$ and $H_x > 0$. If the true group is random but the detector output is nonrandom ($H_y = 0$), then the proficiency is well-defined and equal to zero. On the other hand, if the true group is nonrandom while the detector output is random ($H_x = 0$ and $H_y > 0$), then the proficiency is undefined by the formula I/H_x. However, in the limit as $P_G \to 0$, both $I \to 0$ and $H_x \to 0$ in such a way that the proficiency converges to zero. Therefore, we define the proficiency to be zero if $H_x = 0$ and $H_y > 0$. In the extreme case where both x and y are nonrandom ($H_x = H_y = 0$), the proficiency is defined to be 1.

Procifiency has a nice interpretation using ROC (Reciever Operating Characteristic) curves. Given a specified input $\mathrm{SNR}_{\mathrm{in}}$, each value of precision corresponds to a specific curve. Figure 1 depicts the ROC curves corresponding to $\mathrm{SNR}_{\mathrm{in}} = 0.002$. The box in the upper left corner contains all operating points that achieve a recall of at least 0.75 and a precision of at least 0.50. To achieve this level of performance, we see that the proficiency must be at least 56%.

The complement of proficiency is the *deficiency metric* $\beta = 1 - \alpha$. Table 1 compares the deficiency with the overall error rate (as well as the false-positive and false-negative rates, the precision, and the output signal-to-noise ratio) for different group membership probabilities P_G.

The first two rows in this table show that the error rate is not a sufficient performance metric when group prevalence becomes small. The first row in the table has a group preva-

Table 1: Comparison of Deficiency with Error Rate and other metrics.

Group Prevalence (P_G)	Input SNR ($\frac{P_G}{1-P_G}$)	Miss Rate (F_n)	False-Alarm Rate (F_p)	Deficiency Metric (β)	Error Rate (P_E)	Precision Metric (PV^+)	Output SNR ($G_1 \mathrm{SNR}_{\mathrm{in}}$)
0.1	0.111	0.0001	0.0001	0.0026	0.0001	0.999	999
0.0001	0.0001	0.0001	0.0001	0.136	0.0001	0.5	1
0.0001	0.0001	0.5	0.0001	0.627	0.00015	0.333	0.5

lence of $P_G = 0.1$, while the second row has the much smaller $P_G = 0.0001$. Since the miss and false-alarm rates are both 0.0001, the error rate is 0.0001 for both rows. In contrast, the deficiency is much larger (0.136) in the second row than in the first (0.0026). This is explained by either the precision or the signal-to-noise ratio (one may be mapped into the other). The precision drops from 0.999 in the first row to 0.5 in the second, while the signal-to-ratio drops from 999 to 1.

The last row of the table shows that a dramatic increase in the miss rate (from 0.0001 to 0.5) produces only a small increase in the error rate (0.0001 to 0.00015). In contrast, the deficiency metric increases from 0.136 to 0.627, which flags the fact that half of all group members are expected to go undetected.

4 Illustration

To illustrate the use of the proficiency metric, we show the results of running the group-detection algorithm K-GROUPS [6] on synthetic link data in Monte Carlo simulations of biomolecular experiments. We simulated five target groups of interacting proteins. Each group contained twenty proteins, while the simulated universe contained 10,000 proteins in all. Thus, 9,900 of the proteins were orphans in this simulation, and the $\mathrm{SNR}_{\mathrm{in}}$ for each group was 20/(10000-20) = 0.002. Each simulated experiment was designed to detect from two to six of the proteins in a particular target group. The simulated experimental results contained both clutter and noise. Because each experiment provided only noisy partial information about a single group, multiple experiments were needed to detect the members of all five groups. Our objective was to study the relationship between the number of experiments performed and the resulting group-finding and orphan-finding proficiencies acheived by the K-GROUPS algorithm.

Each experiment produced one link vector of proteins: a putative subset of members from a single target group. However, these links were subject to clutter and noise. In particular, ten percent of the experiments produced bogus links: they identified proteins that were randomly sampled from the universe without regard to the target groups of interest. The remaining ninety percent of experiments produced valid, but noisy, links. Each valid link contained from two to six proteins belonging to one target group, but ten percent of these putative group members were actually noise and did not belong to the group targeted by the experiment. The total number of experiments was the sample size (the number of link vectors); it ranged from 2 to 512. Approximately equal numbers of experiments were used to probe each of the target groups. Table 2 summarizes the parameters of the Monte Carlo simulations.

The resulting mean proficiencies for detecting target groups and orphans are plotted in Figure 2 as a function of sample size (total number of experiments). The ellipses indicate the variability of the proficiencies over the one hundred simulated data sets for each sample size. We see that at least 256 experiments are required to acheive a mean group-finding

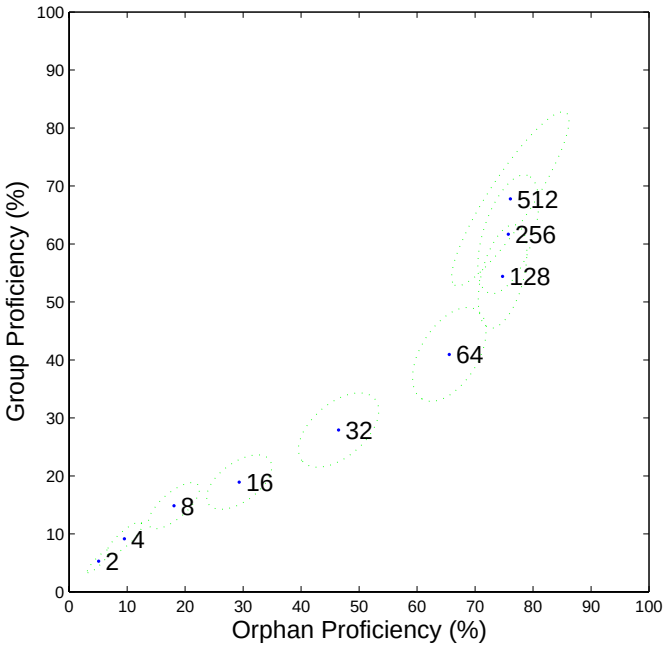

Figure 2: Monte Carlo simulation results. Mean group-finding proficiency and mean orphan-finding proficiency for different sample sizes. Ellipses indicate the amount of variability observed in 100 Monte Carlo simulations of each experiment.

proficiency of 60%, and the mean group proficiency typically increases 5% to 10% for each doubling in sample size. We also note the overlap between the mean proficiencies for sample sizes of 128, 256, and 512. (Histograms of the group-finding proficiencies are displayed in Fig. 3.) These conclusions apply to the simulated world defined by the parameters in Table 2. The results would be different for other choices of the parameters. Our intent here is merely to illustrate the use of the proficiency metric, not to provide an evaluation of K-GROUPS.

Table 2: Parameter values for Monte Carlo simulations.

Parameter	**Value**
# proteins per data set	10,000
# groups per data set	5
# proteins/group	20
# links per data set (sample size)	2, 4, 8, 16, 32, 64, 126, 256, 512
# data sets per sample size	100
# K-GROUPS iterations per data set	10
Link sizes (equally probable)	2, 3, 4, 5, 6
Per-group SNR_{in}	0.0011
Probability of clutter links	0.1
Probability of noise in group-generated links	0.1

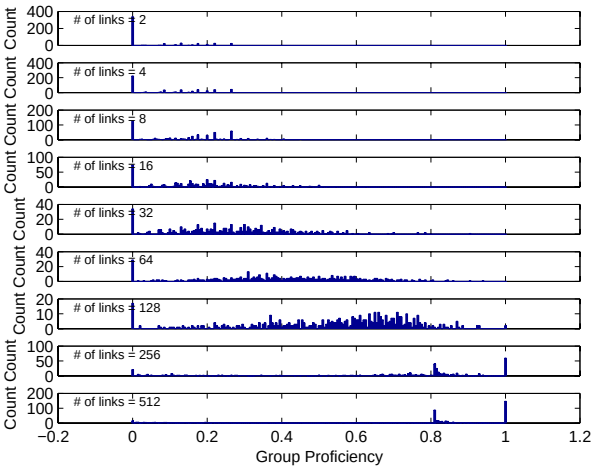

Figure 3: Histograms of Monte Carlo simulation results. Group-finding proficiency for different numbers of links (number of experiments, or sample size).

5 Summary

In this report we define the group-detection problem with reference to the probabilistic generative data model on which CMU's GDA algorithm is based.

We also discuss classical performance metrics for evaluating any group-detection algorithm. The principal conclusions of this discussion are:

1. The probabilistic performance of a detection system, consisting of a link-data source and a group-detection algorithm, "lives" in a three-dimensional space. One coordinate system for this space has the coordinates (1) input signal-to-noise ratio, (2) true-positive rate, and (3) false-positive rate. An alternative set of coordinates is (1) prior probability of group membership for an entity, (1) false-positive rate, and (2) false-negative rate. The choice of coordinate system reflects the preference of the analyst.

2. The classical metrics become insensitive to either the false-positive rate F_p or the false-negative rate F_n at very low input signal-to-noise ratios (when the size of the entity population is much larger than a group of interest). Two classical metrics should be considered simultaneously when evaluating algorithm performance.

3. To avoid the need to consider two metrics simultaneously, a nonclassical metric called proficiency is defined using information theory. Precision is I/H, where I is the mutual information between the true group-membership states of entities and the corresponding outputs of the group-detection algorithm, and H is the entropy of the group-membership states. I is the amount of information that the detector outputs provide about the true group-membership states, while H is the amount of information the detector would have to provide if it were to achieve ideal error-free performance.

4. We recommend that proficiency be used to rank the probabilistic performance of group-finding systems. Proficiency has the advantage that it remains sensitive to both F_p and F_n at low signal-to-noise ratios. It is a neutral metric that evaluates both types of errors while simultaneously accounting for group size relative to the total number of entities.

Finally, we illustrate the application of the proficiency metric: we studied the performance of K-GROUPS in Monte Carlo simulations to determine the effects of sample size in detecting groups of interacting proteins. A single plot conveys simultaneously the mean group-finding proficiency and the mean orphan-finding proficiency as a function of link size.

References

[1] Carpaneto and Toth, "Algorithm 548: Solution of the assignment problem [H]", *ACM Transactions on Mathematical Software*, 6(1):104-111, 1980.

[2] Cover, T. and Thomas, J., *Elements of Information Theory,* John Wiley & Sons, New York, 1991.

[3] H.A. Kautz, B. Selman, and M.A. Shah, "The Hidden Web," *AI Magazine,* 18(2):27–36, 1997.

[4] M.E.J. Newman, "Who Is the Best Connected Scientist? a study of scientific coauthorship networks," *Phys. Rev.,* 64(016131,016132), 2001.

[5] B. Taskar, E. Segal, and D. Koller, "Probabilistic Clustering in Relational Data," *Seventeenth International Joint Conference on Artificial Intelligence,* pp. 870–876, Seattle, Washington, Aug. 2001.

[6] J. Kubica, A. Moore, J. Schneider, Y. Yang, "Stochastic Link and Group Detection," *AAAI*, pp. 798–804, ACM Press, July 2002.

[7] J. Kubica, A. Moore, D. Cohn, J. Schneider, "Finding Underlying Connections: A Fast Method for Link Analysis and Collaboration Queries," International Conference on Machine Learning, in *ICML 2003*, pp. 392–399, AAAI Press, 2003.

[8] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *J. Machine Learning Res."* 3, 2003.

[9] J. Kubica, A. Moore, J. Schneider, "K-groups: Tractable Group Detection on Large Link Data Sets," technical report CMU-RI-TR-03-32, Carnegie Mellon University, Pittsburgh, PA, http://www.autonlab.org, 2003.

A Symbols

The table of symbols appears on the next page.

B Contour Plots

This section contains contour plots, which depict contours of constant error-rate or proficiency for different group signal-to-noise ratios as a function of the false-negative and false-positive rates. The plots appear on pages 12 to 15.

Table 3: Mathematical notation.

Symbol	Meaning
α	Proficiency metric
β	Deficiency metric
A	Detection algorithm
e_k	k-th entity
F_n	False-negative rate, miss rate
F_p	False-positive rate, false-alarm probability
g	Group indexing variable
G_0, G_1	Bayes factors
H_x	Entropy of indicator variable x
H_y	Entropy of indicator variable y
I	Mutual information between x and y
$\mathcal{L}$	Parametric family of random variables (link vectors)
$\mathbf{L}$	Sample of link vectors generated by $\mathcal{L}$
$\log_2$	Base-2 logarithm
n	Number of entities in a link vector
N_E	Total number of entities
N_G	Total number of groups
O_0^-,O_0^+	Prior and posterior odds favoring $x = 1$
O_1^-,O_1^+	Prior and posterior odds favoring $x = 1$
P_b	Probability of a link being group-generated
P_E	Error rate, probability of error
P_G	Probability of entity belonging to a particular group
$\mathbf{P}_x$,$\mathbf{P}_y$	Probability distributions of x and y
$\mathbf{P}_{xy}$	Matrix containing joint probability distribution of x and y
$P_{x\|y}$	Conditional probability of x given y
PV^+,PV^-	Positive and negative predictive values
q	Number groups specified by detection algorithm A
$\mathrm{SNR}_{\mathrm{in}}$	Input signal-to-noise ratio
$\mathrm{SNR}_{\mathrm{out}}$	Output signal-to-noise ratio
T_p	True-positive rate, recall metric
T_n	True-negative rate
x	Indicator variable for a generic entity being in a particular group
$\mathbf{x}$	Vector of indicator variables for group membership
$\mathbf{x}_g$	Vector of indicator variables for membership in group g
$\mathcal{X}$	Random variable modeling group membership
y	Indicator variable for the detection-algorithm output
$\mathbf{y}$	Vector of indicator variables for a detected group
$\mathbf{y}_g$	Vector of indicator variables for detected group g

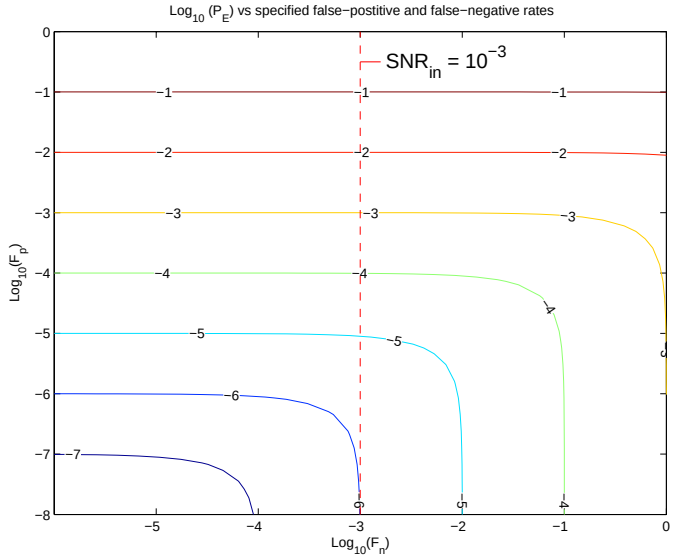

Figure 4: Contour plot of log error-rates for $\mathrm{SNR}_{\mathrm{in}} = 10^{-3}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

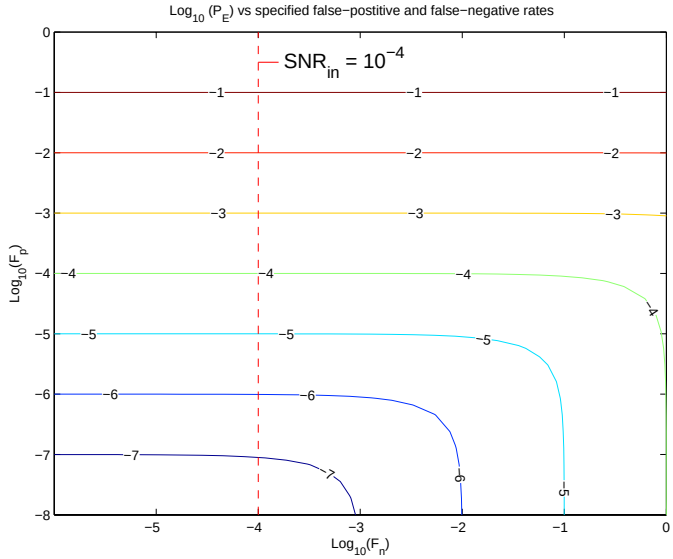

Figure 5: Contour plot of log error-rates for $\mathrm{SNR}_{\mathrm{in}} = 10^{-4}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

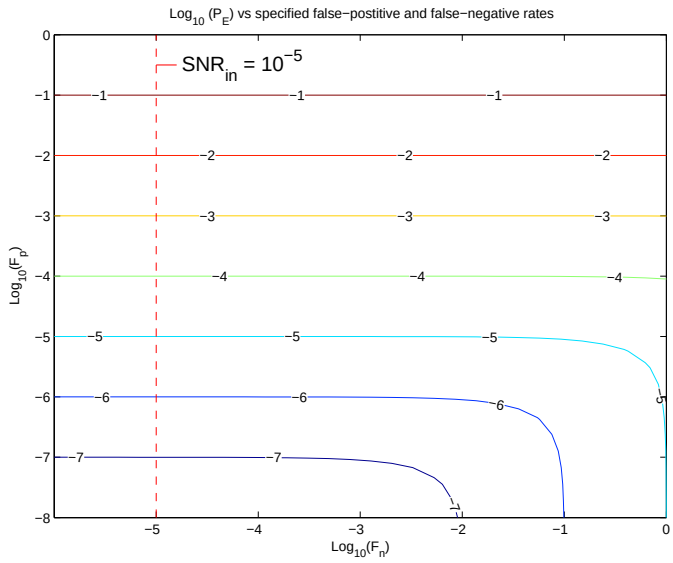

Figure 6: Contour plot of log error-rates for $\mathrm{SNR}_{\mathrm{in}} = 10^{-5}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

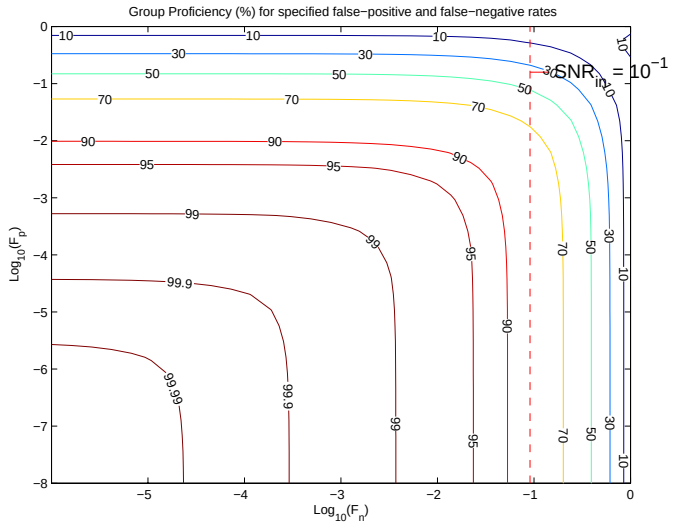

Figure 7: Contour plot of group-finding proficiency for $\mathrm{SNR}_{\mathrm{in}} = 10^{-1}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

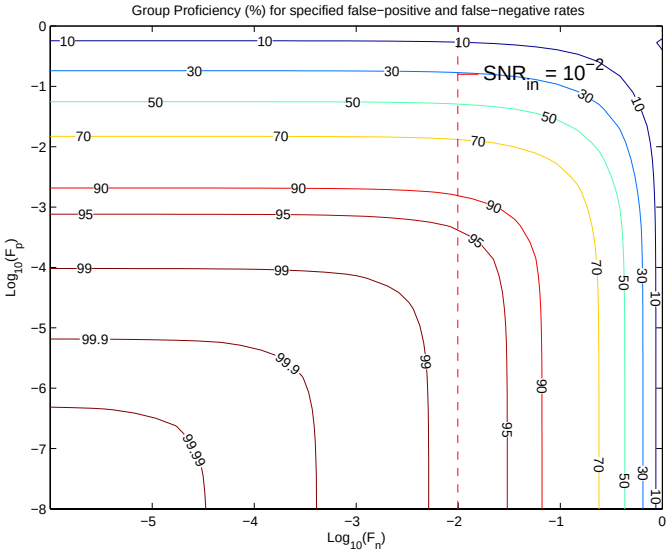

Figure 8: Contour plot of group-finding proficiency for $\mathrm{SNR}_{\mathrm{in}} = 10^{-2}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

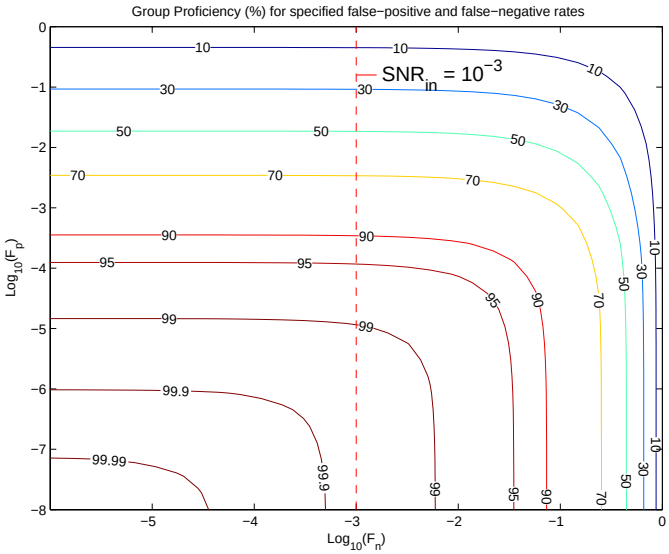

Figure 9: Contour plot of group-finding proficiency for $\mathrm{SNR}_{\mathrm{in}} = 10^{-3}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

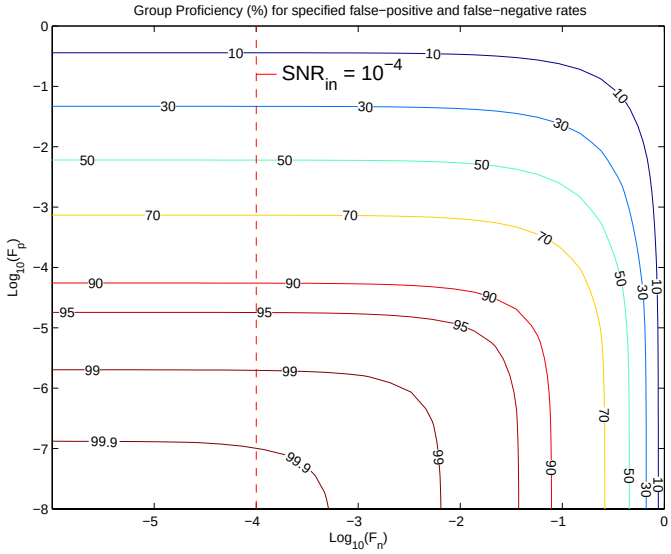

Figure 10: Contour plot of group-finding proficiency for $\mathrm{SNR}_{\mathrm{in}} = 10^{-4}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.

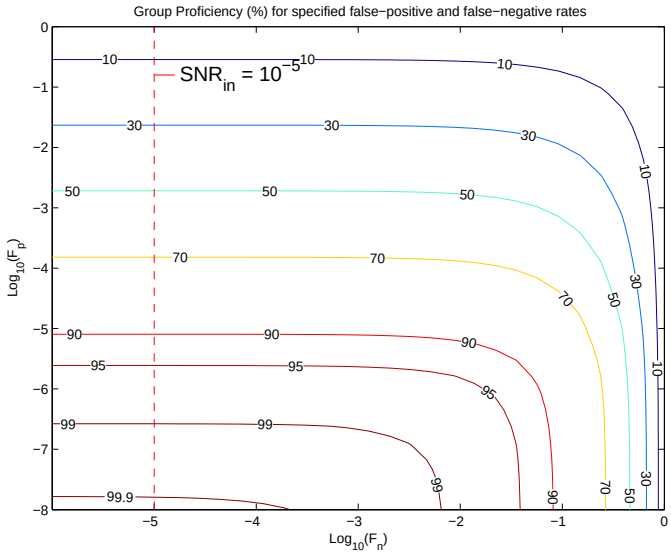

Figure 11: Contour plot of group-finding proficiency for $\mathrm{SNR}_{\mathrm{in}} = 10^{-5}$. Dashed line marks the upper bound on false-negative rate for this $\mathrm{SNR}_{\mathrm{in}}$.