MATHEMATICAL BACKGROUND OF PERSONALIZED NETWORK MEDICINE

PNM replaces the analysis of data values by the analysis of relationship patterns between these data values. The use of graphs results in a compact representation of these relationship patterns, which would become prohibitively complicated if we would attempt to quantitatively capture the same information in another way (imagine, for example, statistical models where all independent variables are covariates). Nevertheless, the use of graphs for capturing the high-level data relationship information still does not solve the problem how to simplify the processing of very complicated input. Even in small space of relationships between just 5 variables, we found that 2631 patients were characterized by 166 different clinical profiles PRP_i. Out of those, 117 (70.5% majority) were found for just five or less patients and, in other extreme, the two most common personal clinical profiles were shared only by 197 and 217 patients, respectively. From this perspective, despite of compact mathematical form of network-captured relationship information, most individuals in the study appeared as "outliers", because of the inherent complexity of the unprocessed relationship pattern data.

Therefore, additional mathematical step is needed before meaningful data analysis and interpretation of the higher-level information, encoded in the data relationship patterns, is possible. The conceptual background for this additional step, which defines the PNM, is simple: we take into consideration that relationships between data in clinical studies are generally not random, but structured. For example, in concrete application, the probability that a person from a family with cleft history will have lip whorl is higher than random chance of such relationship. These "transition probabilities" between the clinical variable values and phenotype were (or might have been) established as "associations" by separate studies, but we need to integrate them into the full context of other personal descriptors of an individual. Conceptually, this equals to fundamental paradigm change. Instead of looking for biases in data values, capturing i.i.d. responses of subject systems to phenotype condition(s), we want to decode information, underlying the non-trivial, function-related relationships between the observed data values, while all other "parasitic", non-informative relationships are considered in maximally unbiased way. With this adjustment of the analysis goals, the measured variable values represent the signals, and observed co-occurrence frequencies are transition probabilities between states, generating these signals. Formalism of information entropy characterization of the study data and relationships between their values, considered as the information source, where the personal context of an individual modulates the observed phenotype-related signals, is the natural choice for solving this problem. Next paragraphs summarize the essentials of the formal integration of various fundamental mathematical results, leading to the PNM implementation.

Graph is defined as the specific set of vertices, edges (connections between vertex pairs) and the topology (set of prescriptions how the vertices are connected by edges. Topology of a graph with ***n*** vertices can be captured by ***n***x***n*** adjacency matrix).

For our purposes, we need to generalize this graph definition and simultaneously put restriction on its topology: **a)** we define study graph ***S*** as ***k***-partite graph with **b)** vertices carrying additionally the vertex potentials and **c)** the relationship information, carried by edges is extended by edge weights (see below for definition of this double-weighting).

Definition of vertices and vertex potentials: To every observed clinical variable, we assign a set of ***v*** vertices. Each of this vertex ***v***-tuple will carry a potential function value, representing the clinically relevant variable value or type. Thus, for gender we will have ***v*** = 2 with the potential function values V_1 = male, V_2 = female. For blood bilirubin variable, we might have ***v*** = 3, with V_1=<0-2 mg/dL = normal, V_2 = 2-5 mg/dL = elevated, and V_3 = >5 mg/dL = high, or other definition of the ranges, as used in the best clinical practice for a given phenotype. While it is generally true that some information is lost by discretization of continuous variables, in PNM we have two mechanisms how to compensate for that “loss”. First, by strict adherence to the best clinical practices for clinical interpretation of individual variable levels, we encode into the vertex potentials the same information that is used in medicine. Second, PNM adds to these discretized values the information about their relationships, which more than compensate for the minimal information loss introduced by the expert-guided discretization.

Definition of edges and edge weights: Edge in ***S*** indicates, that a co-occurrence of specific values of a variable pair was observed in a study. Edge weight is defined as co-occurrence frequency for every variable level pair. With this definition, the edge weigts are the estimates of the transition probabilities between the variable states.

Restriction to *k*-partite topology of *S* is the direct consequence of the above definitions: Edges represents primarily the observed relationships between the clinical variable observations for a person. Thus, if we consider person’s weight, the subject cannot be simultaneously obese and underweight. Therefore, in clinical data, we will never study the value of a variable within the context of another value of the same variable for the same person. Therefore there will never be edges between vertices representing the observable ranges of the same variable, which by definition lead to ***k***-partite graph ***S***.

With the above definition of graph ***S***, the relationship patterns, experimentally observed in the study data are represented as (clinical variable) states and transition probabilities. Now the question is, if any internal structure in such data relationship set exists, which would provide a natural simplification of its complexity. There is unique answer to this question, which can be found by analysis of information entropy H. The fact, that H = $-p_i\ln(p_i)$ reflects very fundamental inherent property of any system with the structured information: in such a system, there are many relationship structures that contribute identically to the total information entropy.

The logarithmic measure $-p_i\ln(p_i)$ is unique function that assigns correctly the identical fraction of entropy to all those equivalent information structures (see the explanatory scheme on the next page). This uniqueness makes the entropy function the only available tool for quantitative labeling of all equivalent relationship patterns, which we want to analyze. In our example, if we could assign entropy to all 166 types of clinical profiles *PRP$_i$* found in the study, those with identical H are carrying equivalent information about the patients, could be grouped into one clinically equivalent class, and the desired simplification of the information structure would be achieved.

General solution of the entropy characterization of the information, captured by the relationship graphs can be found by applying the maximal entropy principle to graph sets. The solution is defined by the following theorem: Let G be a finite set of graphs with elements PRP_i. Distinguish an element HL_j of G. Define $\delta(PRP_i, HL_j)$ as a distance metric on G. Define $p(PRP_i)$ as probability of finding PRP_i in G. The probability distribution that maximizes the entropy H(*p*), subject to constraints that

$$\sum_{PRP_i \in G} \delta(PRP_i, HL_j)\, p(PRP_i) = \nu$$

satisfies

$$p(PRP_i) = c(HL_j, \tau)e^{-\tau\delta(PRP_i, HL_j)},$$

where

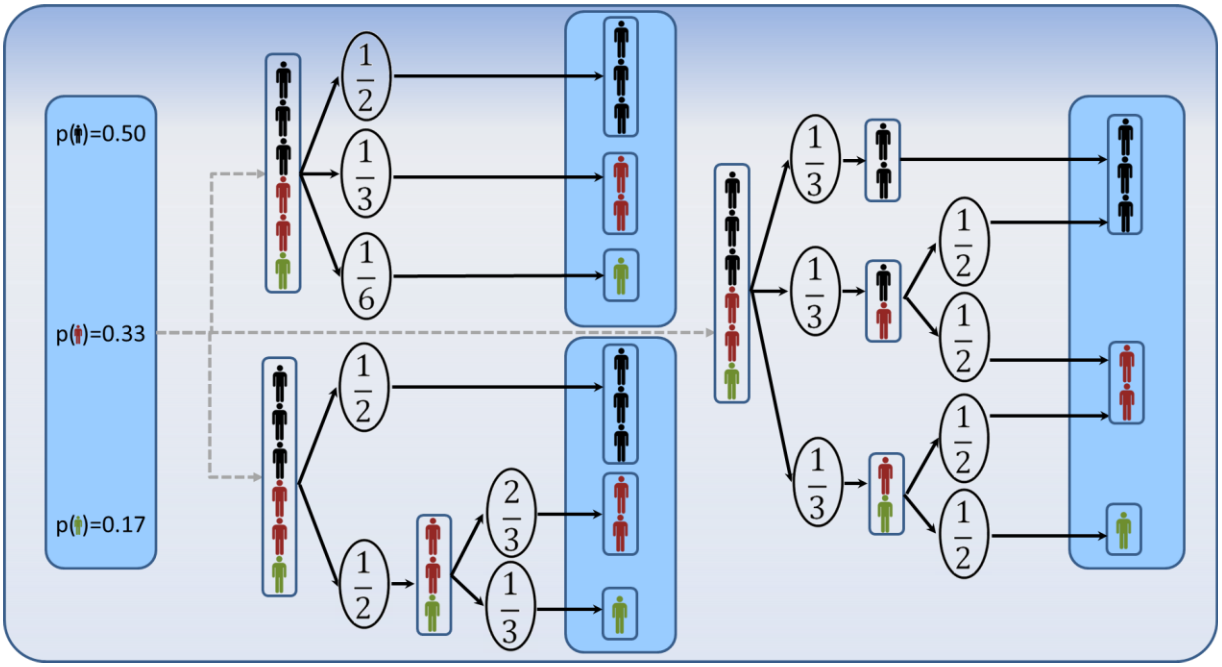
$$c(HL_j, \tau)^{-1} = \sum_{PRP_i \in G} e^{-\tau\delta(PRP_i, HL_j)}$$

and τ is the unique solution to

$$\frac{d\ln c(HL_j, \tau)}{d\tau} = \nu.$$

In short, for structured clinical information, captured by our graphs, the solution of the problem of finding the correct (simplest) internal information structure requires quantitative measure $\delta(PRP_i, HL_j)$ of graph–graph distances relatively to HL_j. From condition that both PRP_i and HL_j are elements of G and by defining PRP_i as ***k***-partite graph, it follows, that HL_j is also ***k***-partite graph. Choice of $\delta(PRP_i, HL_j)$ is restricted only by the conditions imposed on the set metrics (positive definiteness, transitiveness and triangular inequality). Our choice of $\delta(PRP_i, HL_j)$ as edit distance of PRP_i and HL_j was guided by several advantages, listed below. We are investigating properties of other definitions of $\delta(PRP_i, HL_j)$ which are more suitable for other specialized applications.

Edit distance was introduced by Hamming



in the context of error correcting codes. It is defined as the number of edge discrepancies between PRP_i and HL_j. It can be shown, that for a graph with ***m*** vertices, the geometric representation of its edit distance from another graph is an $r = \binom{m}{2}$ dimensional hypercube. For our purposes, important property of this representation is the symmetry of the hypercube, from which it follows that $c\left(HL_j, \tau\right)^{-1}$, (which is the graph theoretical equivalent of the ensemble partition function of statistical thermodynamics) is independent of HL_j. This fundamental result has two consequences. First, it nullifies all (uniformed) critics of PNM, who try to imply, that its result are arbitrary, because of liberty in choosing HL_j. Second, because of this symmetry, the analytical solution of the entropy maximization equations can be derived (#). The analytical form of the graph set partition function $c\left(HL_j, \tau\right)^{-1}$ is

$$\sum_{PRP_i \in G} e^{-\tau\delta\left(PRP_i, HL_j\right)} = \sum_{k=0}^{r} \binom{r}{k} e^{-\tau k} = (1 + e^{-\tau})^r = c\left(HL_j, \tau\right)^{-1}$$

The analytical formula for the Lagrange multiplier τ, maximizing the entropy H(*p*) is obtained by differentiation of the above partition function:

$$\tau = \frac{(rn)^{-1} \sum_{i=1}^{n} \delta\left(PRP_i, HL_j\right)}{1 - (rn)^{-1} \sum_{i=1}^{n} \delta\left(PRP_i, HL_j\right)}$$

With these results, we can finally derive the analytical form of the probability distribution of the $\delta\left(PRP_i, HL_j\right)$ edit distance values 0,1,…,x,x+1,…,k-1,***k***

$$p(\delta = x) = \binom{r}{x} (1 + e^{-\tau})^{-r} e^{-x\tau}$$

Plots of these functions are shown in **Figs. A1** and **A2**.

Fig. A3 summarizes these results and provide the blueprint of the understanding of PNM processed information and the roadmap for the implementation in the concrete applications. The mean values of distribution of the observed personal distances $\delta\langle PRP_i, HL_j\rangle$, computed for the cohort or various sub-cohorts (the x-axis

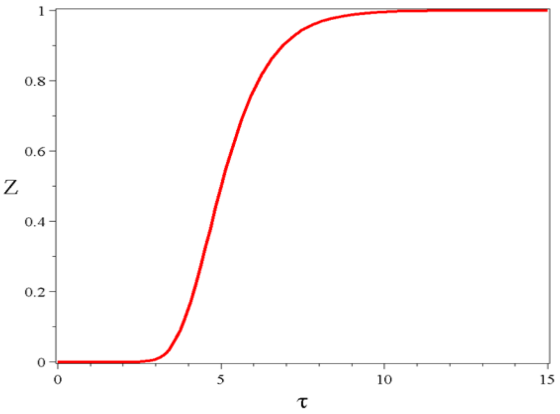


Figure A1. Plot of the graph partition function $c\left(HL_j, \tau\right)^{-1} = (1 + e^{-\tau})^r$

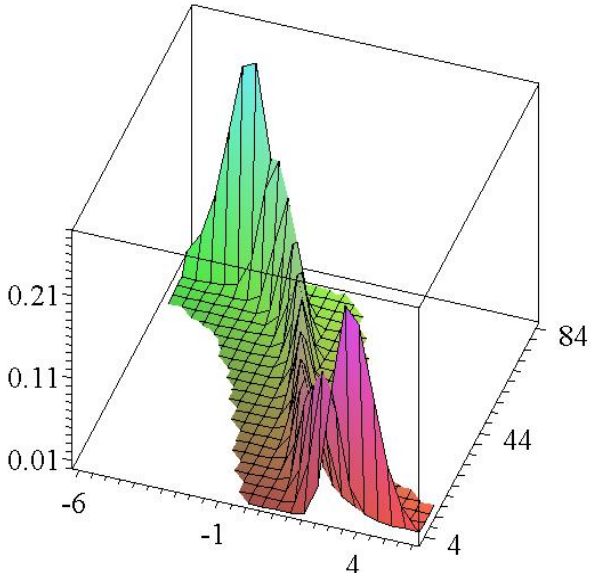


Figure A2. Plot of the probability distributions $p(\delta = x) = \binom{r}{x} (1 + e^{-\tau})^{-r} e^{-x\tau}$ for various r and τ

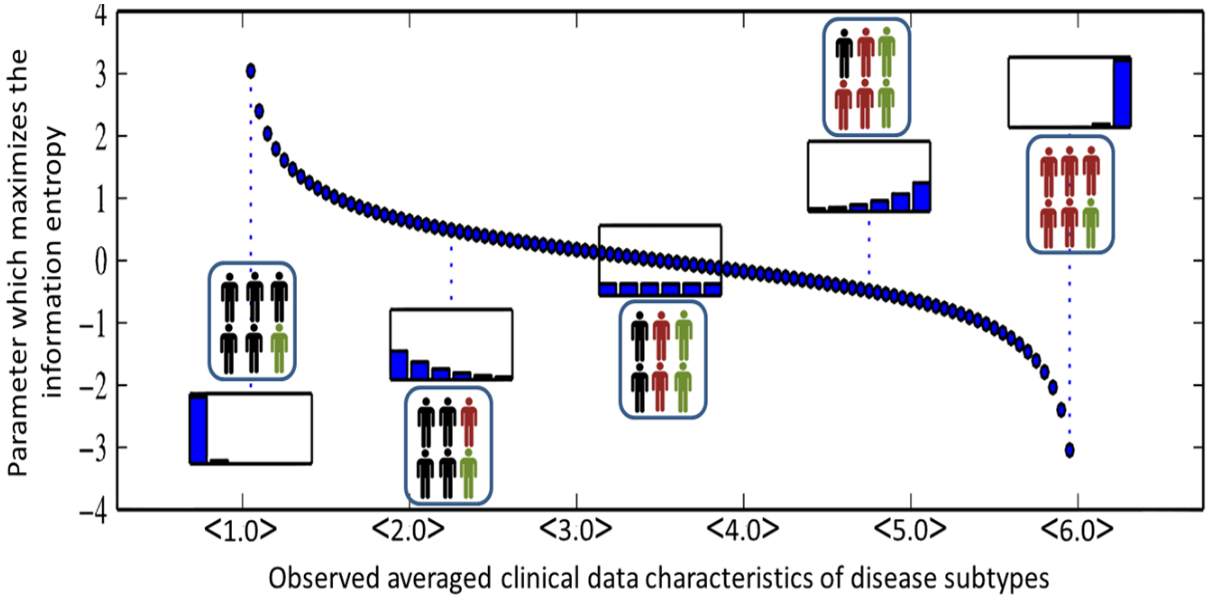


Figure A3. Schematic explanation of the relationship between the values of actual mean values of the distances $\delta\left(PRP_i, HL_j\right)$ (x-axis), Lagrange multiplier τ (blue circles, y-axis), probability $p(\delta = x)$ distributions and clinical heterogeneity of the studied cohort (different colors of the personal icons).

in the Fig.A3) are linked by the analytical formulae for graph set partition function $c\left(HL_j, \tau\right)^{-1}$ and Lagrange multiplier τ to the concrete distributions of the distances $\delta\left(PRP_i, HL_j\right)$. The extreme values of the personal distances lead to the probability distributions that are biased towards similarity to some (and dissimilarity to other) landmarks HL_j. This is the mechanism extracting optimally the structured information for one “phenotype dimension”, characterized by a clinical status represented by a given heterogeneity landmark. As the dimensionality of the information is defined by the number of informative HL_j's, we have similar relationships for all other HL_j's.

Dealing with multidimensionality is simple, because entropy (as a state function) is additive. This allows for probing of the complete phenotype space in all projections of the overall data into respective informative HL_j dimensions. Use of these independent projections was the rationale for concrete implementation of this strategy through systematic testing of Euclidean distance distributions for all pairs of HL_j's,.