The Kingdom-Capacity of a Graph: On the Difficulty of Learning a Graph Labeling

K. Pelckmans[1], J.A.K. Suykens[1], and B. De Moor[1]

K.U. Leuven, ESAT, SCD/SISTA, Kasteelpark Arenberg 10, Leuven, B-3001, Belgium
e-mail: kristiaan.pelckmans@esat.kuleuven.be
WWW home page: http://www.esat.kuleuven.ac.be/scd/

Abstract. This short paper[1] establishes a measure of capacity of a graph which can be used to characterize the complexity of learning labels of a graph.[2] We denote it as the Kingdom-capacity (K-capacity) of a graph. It is shown how this notion is implied by the definition of an out-of-sample extension rule predicting the labels of unobserved nodes. We proceed by proposing an efficient way to compute the K-capacity of a given graph based on the analogy with a simple strategy game. It is shown how this measure of capacity can be used to construct a nontrivial generalization bound for transductive inference of the label of a graph node.

1 INTRODUCTION

The aim of learning over graphs received increasing attention in recent years, see e.g. [2]. Such tasks can often be abstracted in terms of associating a $+1$ or -1 label to the respective nodes. Such problems can be related to the task of looking for a minimal cut and variations, see e.g. [8]. Capacity concepts are omnipresent in graph theory, they mostly focus on issues of connectedness or are based on maximal matching or coloring results. The Shannon capacity for example characterizes the maximal information which can be transmitted through a network. An efficient SDP relaxation was devised, see e.g. [5]. This work aims at a similar result in the context of learning over graphs.

Herefor, a capacity measure is defined and analyzed, characterizing the complexity of learning the labels of a graph. This notion follows the intuition of maximal margin classifiers, and follows the same line of thought as the classical notion of the shattering number and the VC dimension in learning theory [4, 9]. The Kingdom-Capacity (K-capacity) expresses how many nodes can chose their own label at will ('are king to a neighborhood'). Although this capacity is combinatorial in nature (and hence difficult to compute), it is indicated how a nontrivial upper- and lowerbound respectively can be computed efficiently. More specifically, linear programming problems are formulated yielding such a bound. The formulation of those problems is explained by analogy to a simple strategy game in which one determines the maximal number of king nodes who can all defend their territory (potentially) succesfully.

[1] (KP): BOF PDM/05/161, FWO grant V4.090.05N; - (JS) is an associate professor and (BDM) is a full professor at K.U.Leuven Belgium, respectively. (SCD:) GOA AMBioRICS, CoE EF/05/006, (FWO): G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, (ICCoS, ANMMM, MLDM); (IWT): GBOU (McKnow), Eureka-Flite2 IUAP P5/22,PODO-II,FP5-Quprodis; ERNSI;

[2] We like to thank professor J. and U. for constructive discussions.

The following Section formulates this capacity, and motivates this capacity in learning theory. Section 3 discusses a practical approach to efficiently assess this measure of complexity. Section 4 discusses a numerical example, and reviews some further consequences of this line of thought.

2 GRAPH CUTS AND HYPOTHESIS SET

Let a graph G be defined as a set of nodes $V = (v_1, \ldots, v_n)$ and corresponding edges $\{w_{ij} \geq 0 \text{ between } v_i \text{ and } v_j\}_{i,j=1}^n$. Let the nodes have a corresponding label $\{q_i \in \{-1, 1\}\}_{i=1}^n$. Let the vector of positive weights w_i be defined as $(w_{i1}, \ldots, w_{in})^T \in \mathbb{R}_+^n$. Define the neighborhood function $f(v_*)$ of any node v_* as

$$f(v_*) = \sum_{j=1}^{n} w_{*j} q_j = w_*^T q, \tag{1}$$

where $q = (q_1, \ldots, q_n)^T \in \{-1, 1\}^n$. Then the neighborhood rule becomes '$\mathrm{sign}(f(v_*))$' as suggested e.g. in [2, 3]. Label y_i is defined to be consistent with its prediction rule if

$$q_i f(v_i) = q_i \sum_{j=1}^{n} w_{ij} q_j = q_i (w_i^T q) \geq 0, \tag{2}$$

and consistent with a margin $\rho > 0$ if $q_i(w_i^T q) \geq \rho$. The set of different hypothical labelings $q \in \{-1, 1\}^n$ where each labeling is consistent with corresponding neighborhood rules with a certain margin $\rho > 0$ is thus defined as

$$H_\rho = \left\{ q \in \{-1, 1\}^n \;\middle|\; q_i(w_i^T q) \geq \rho, \;\; \forall i = 1, \ldots, n \right\}. \tag{3}$$

Let $|H_\rho|$ denote the cardinality of the set H_ρ. It is seen intuitively that this mechanism of self-consistency may be used to reduce the number of permitted labels: $|H_\rho|$ can be expected to be smaller than 2^n. Before making this statement more precise, a motivation for this restriction of the hypothesis set is given for the case of transductive learning graph labels.

2.1 TRANSDUCTIVE INFERENCE ON GRAPHS

We show how the rate of learning of labels of the deterministic graph G can be expressed in terms of (a bound on) the cardinality of this set. Therefor, the setting of transductive learning is adopted: let a graph $G = (w, v)$ with n nodes $\{v_1, \ldots, v_n\}$ and connections $w = \{w_1, \ldots, w_n\}$ be fixed beforehand. Let to each node a deterministic label $y_i \in \{-1, 1\}$ be associated. The random mechanism becomes apparent through the choice of which nodes are seen during the training phase. Let the random variable S_i for an $1 \leq i \leq n$ denote a randomly chosen node, governed by a fixed but unknown distribution function F_S. Then we study sets of samples $S = \{S_1, \ldots, S_{n_y}\}$ which does not contain a node twice or more *(sampled without replacement)*. The algorithm observes the graph G and the labels of a subset of nodes y_S defined as $y_S = \{y_{S_j}\}_{j=1}^{n_y}$ where $S \subset \{1, \ldots, n\}$ and $n_y < n$. We observe a specific sample $y_{\hat{S}} = \{y_{\hat{S}_1}, \ldots, y_{\hat{S}_{n_y}}\}$. Now, a crucial assumption is that this

sample S is taken i.i.d. from the nodes. Formally, let the actual classification risk be defined as

$$\mathcal{R}(F_S; q) = \frac{1}{n_y} \int \sum_{i \in s} I(y_i w_i^T q < 0)\, dF_S(s), \tag{4}$$

where the inicqtor function $I(z < 0)$ is defined as one if $z < 0$ and zero otherwise. The empirical counterpart becomes

$$\hat{\mathcal{R}}(\hat{S}; q) = \frac{1}{n_y} \sum_{i \in \hat{S}} I(y_i w_i^T q < 0). \tag{5}$$

Application of Serfling's tail inequality as in [7, 6] gives the following result.

Theorem 1 (Complexity of Learning) *With probability $0 < 1 - \delta < 1$, the following inequality holds for all labelings $q \in H_\rho$:*

$$\mathcal{R}(F_S; q) \leq \hat{\mathcal{R}}(\hat{S}; q) + \sqrt{\left(\frac{n - n_y + 1}{n}\right) \frac{\ln |H_\rho| - \ln \delta}{2 n_y}}. \tag{6}$$

Proof: Since only a finite number of hypothesis can be chosen as stated previously, the bound follows directly from application of the Union bound on Serfling's inequality, see e.g. [9], Ch. 4. ■

It becomes apparent that a good uppebound to $|H_\rho|$ is important for the usefullness of the bound. This theorem yields the formal motivation for the following statements *the expected loss of using the rule '*$\mathrm{sign}(w^T q)$*' on n_y randomly sampled nodes would not be too different from the loss observed*. The learning problem computing a hypothesis $q \in H_\rho$ which corresponds maximally with the n_y provided labels becomes

$$\min_{q \in \{-1,1\}^n} \sum_{i \in \hat{S}} I(y_i q_i < 0) \quad \text{s.t.} \quad q \in H_\rho. \tag{7}$$

Since the predictor rule '$\mathrm{sign}(w^T q)$' is to be constraint to be consistent on all nodes with the labels (see definition of H_ρ in (3)), this rule can be used to predict the label of nodes which are not in $\hat{S}$. The study of an efficient algorithm to solve this learning problem, and a formal relation to approaches based on a minimal cut is the topic of a forthcoming paper.

2.2 KINGDOM CAPACITY

We now advance to a more complex way to characterize the complexity of the class H_ρ. The Kingdom capacity (K-capacity) - denoted as $\vartheta(\rho)$ - is defined as the maximal number of nodes which can chose their label at will ('king') such that remaining nodes ('vassals') can be labelled making the king nodes consistent with theirselves. Let q_s be defined as the restriction of a vector $q \in \{-1, 1\}^n$ to the set of indices $s \in \{1, \ldots, n\}$. Let $q_{\setminus s}$ denote the difference set $q \setminus q_s$ ($q_{\setminus s}$ is also denoted as the completion of q).

Definition 1 (Kingdom Capacity) *Given a fixed graph $G = (v, w)$, the K-capacity $\vartheta(\rho)$ is defined as follows*

$$\vartheta(\rho) = \max_{s \subset \{1,\ldots,n\}} |s| \quad \textit{s.t.} \quad \forall q_s \in \{-1,1\}^{|s|}, \;\; \exists q_{\setminus s} \in \{-1,1\}^{n-|s|} \;\Big|\; q_i (w_i^T q) \geq \rho \quad \forall i \in s. \tag{8}$$

This definition is motivated by the following bound on the cardinality of the set H_ρ.

Theorem 2 (Bound to Cardinality) *For any graph G with n nodes and a K-capacity of $\vartheta(\rho)$, the hypothesis class H_ρ contains at most a finite number of hypotheses. This number is bounded as follows*

$$|H_\rho| \leq \binom{n}{\vartheta(\rho)} 2^{\vartheta(\rho)}. \tag{9}$$

Proof: At first, consider a fixed graph G with n nodes and corresponding connections $\{w_{ij} \geq 0\}_{i \neq j}$. Let s be a subset of $\{1, \ldots, n\}$ of *maximal cardinality* such that for every assignment of $+1$ either -1 to the nodes q_s, one can find a completion $q_{\setminus s} \in \{-1, 1\}^{n-|s|}$ such that for the nodes in s the rule $q_i(w_i^T q) \geq \rho$ holds. Formally

$$s \subset \{1, \ldots, n\}: \;\; \forall q_s \in \{-1, 1\}^{|s|}, \exists q_{\setminus s} \in \{-1, 1\}^{n-|s|}: \;\; q_i(w_i q) \geq \rho, \; \forall i \in s. \tag{10}$$

Remark that there may be more than one such sets $s \subset \{1, \ldots, n\}$ of the same (maximal) cardinality. More precisely, there may be maximally $\binom{n}{|s|}$ such a sets. This definition implies that the graph G permits at most $\binom{n}{|s|} 2^{|s|}$ different hypotheses. Indeed, assume that $|H_\rho| > \binom{n}{|s|} 2^{|s|}$. As q can only take two different values, there is at least one subset $s' \subset \{1, \ldots, n\}$ with higher cardinality ($|s| < |s'|$) such that all $2^{|s'|}$ possible combinations $q_{s'} \in \{-1, 1\}^{|s'|}$ satisfy the selfconsistency, contradicting the assumption of maximality.

By definition, the K-capacity equals the cardinality of this maximal set s, hence we denote such a set as $s_{\vartheta(\rho)}$. The above reasoning yields the following inequality

$$|H_\rho| \leq \binom{n}{\vartheta(\rho)} 2^{\vartheta(\rho)}. \tag{11}$$

Since we do not restrict the maximal number of sets $s_{\vartheta(\rho)}$, this bound holds for any given graph with n nodes and a K-capacity of $\vartheta(\rho)$, yielding (9). ■

It is seen that this reasoning is similar in spirit to Sauer's lemma and extensions as discussed e.g. in [1].

3 ASSESSING THE K-CAPACITY

Efficient ways to compute an upper- and lowerbound to the K-capacity respectively are outlined.

3.1 UPPER-BOUND TO THE K-CAPACITY: THRESHOLDING EDGES

A first method gives an intuitive bound on the K-capacity based on neglecting the non-important connections. More precisely, if a certain connection w_{ij} is larger than all other connections $\{w_{ik}\}_{k \neq j}$, (i.e. $w_{ij} \geq \sum_{k \neq j} w_{ik} - \rho$) a cut of the former cannot be corrected by any combination of latter weights. Thus a necessary condition is that no cut may occur between node v_i and node v_j, hence reducing the maximal set of free nodes by a unit.

Proposition 1 (Bound by Thresholding Edges) *Let $\tilde{\vartheta}(\rho)$ be defined as*

$$\tilde{\vartheta}(\rho) \triangleq \max_{s\in\{1,\dots,n\}} |s| \quad \textit{s.t.} \quad \forall i \neq j \in s: \;\; w_{ij} < \sum_{k\neq j} w_{ik} - \rho, \; \textit{AND} \;\; w_{ij} < \sum_{k\neq i} w_{kj} - \rho. \tag{12}$$

Then $\vartheta(\rho) \leq \tilde{\vartheta}(\rho) \leq n$.

Proof: We show that $\vartheta(\rho) \leq \tilde{\vartheta}(\rho)$. Suppose $w_{ij} \geq \sum_{k\neq j} w_{ik} - \rho$ and $q_i q_j = -1$ (i.e. cut between v_i and v_j), then the following inequality holds

$$q_i \left(w_{ij} q_j + \sum_{k\neq j} w_{ik} q_k \right) \leq -w_{ij} + \sum_{k\neq j} w_{ik} \leq \rho, \tag{13}$$

hence contradicting the condition of the K-capacity in (8). This implies that no cut can be made between nodes with too strong a connection. Maximizing with this (only) necessary condition gives the uppperbound. ■

To compute this quantity, construct a graph G_ρ with the same n nodes v as previously, and weights w_ρ such that $w_{ij,\rho} = w_{ij}$ if $w_{ij} \geq \sum_{k\neq j} w_{ik} - \rho$ or $w_{ij} \geq \sum_{k\neq i} w_{kj} - \rho$, for all $i \neq j$, or zero otherwise. Then $\tilde{\vartheta}(\rho)$ simply equals the number of disconnected components in G_ρ.

Remark that this notion is related to the classical notion of capacity of a graph defined as follows (see e.g. [5])

$$\text{cap}(G) = \max_{s\in\{1,\dots,n\}} |s| \quad \text{s.t.} \quad \forall i \neq j, \; w_{s_i,s_j} = 0. \tag{14}$$

This approach of thresholding often yields however overly pessimistic estimations ($\tilde{\vartheta}(\rho) \approx n$) of the K-capacity, especially when all non-zero weights take values of similar magnitude.

3.2 UPPER-BOUND TO THE K-CAPACITY: LAZY KINGS

A more thight upper-bound can be obtained from a convex optimization problem as follows. To introduce the methodology, we encode the problem based on the analogy with a simple strategy game. Here, we set a node to a *king* if it can chose its own label freely. The remaining nodes are a *vassal*, as they are to support their kings' will. This is incoded asa binary variable ξ_i for each node v_i: $\xi_i = 1$ means that v_i is a king, $\xi_i = 0$ means that v_i is a vassal. Now it is clear that we look for the maximal number of kings which can be supported by a given graph. The first formulation assumes the kings are lazy: they are happy as long as their are enough vassal neighbours which can be enslaved when needed. This means that they need not be suspicious and vassals are thus not governed by a single king. This idea implements a neccessary condition for the bound to hold, resulting in an upper-bound. Making this idea formal, we obtain that the number of vassals ($\xi_i = 0$) multiplied by their connection weight should majorate the weighted number of kings in a king's vicinity, or for all $i = 1, \dots, n$:

$$v_i\text{king} \Rightarrow \sum_{j=1}^{n} w_{ij}\xi_j \leq \sum_{j=1}^{n} w_{ij}(1-\xi_i) - \rho. \tag{15}$$

Reformulating the left hand side and forcing the constraint to be trivial when $\xi_i = 0$ yields the expressions $2\sum_{j=1}^{n} w_{ij}\xi_j \leq (2-\xi_i)\sum_{j=1}^{n} w_{ij} - \rho$ for all $i = 1, \ldots, n$. Combining the above reasoning gives the integer programming problem

$$\max_{\xi_i \in \{0,1\}} \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad 2\sum_{j=1}^{n} w_{ij}\xi_j \leq (2-\xi_i)\sum_{j=1}^{n} w_{ij} - \xi_i\rho, \;\; \forall i = 1, \ldots, n. \tag{16}$$

Relaxating the integer constraints to $\xi_i \in [0, 1]$ for any $i = 1, \ldots, n$ can only increase the maximum, motivating the upperbound

$$\overline{\theta}(\rho) = \max_{\xi_i} \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \begin{cases} 2\sum_{j=1}^{n} w_{ij}\xi_j \leq (2-\xi_i)\sum_{j=1}^{n} w_{ij} - \xi_i\rho & \forall i = 1, \ldots, n. \\ 0 \leq \xi_i \leq 1 \end{cases} \tag{17}$$

Remark that since (16) may only take integer values, the integer part of this value may be considered as the upperbound without loss of generality such that $\lfloor \overline{\theta}(\rho) \rfloor \geq \vartheta(\rho)$. The upperbound is not tight as in typical situations (i.e. choice of label of the king node), a king may not win all its neighboring vassal nodes for its personal sake. At the time of labeling, competition of the vassal nodes will divide the vassal nodes over the kings. The following section uses a fixed assignment of vassals to a unique king to obtain a lowerbound.

3.3 LOWER-BOUND TO THE K-CAPACITY: SUSPICIOUS KINGS

We show how one can compute efficiently a lower-bound to the K-capacity. what is the highest number of suspicious kings who can govern their own disjoint subgraph simultaneously and independently. To formalize this problem, let again the binary variable $\xi_i \in \{0, 1\}$ denote whether a node v_i is a *vassal* ($\xi_i = 0$), or a *king* ($\xi_i = 1$). Then, define for every node v_j a vector of binary variables $\beta_j = (\beta_{j1}, \ldots, \beta_{jn})^T \in \{0, 1\}^n$ encoding to which king node it will be dedicated: exactly one element of β_j is to be one, the others are zero. This property may be encoded as follows

$$\sum_{j \neq i} \beta_{ij} \leq (1 - \xi_i), \;\; \forall i = 1, \ldots, n. \tag{18}$$

Furthermore, a node might be a king ($\xi_i = 0$), if it governs enough neighbors to possess a superiority over the remaining nodes when they would gather against the king (hence the denpminator 'suspicious').

$$v_i \text{ is a king} \Rightarrow \sum_{j=1}^{n} w_{ji}\beta_{ji} \geq \sum_{j=1}^{n} w_{ji}(1 - \beta_{ji}) + \rho. \tag{19}$$

If not, it might thrown in its lot with a neighboring king (vassal), and the above constraint becomes obsolete. This reasoning results in the following integer programming problem

$$\theta(\rho) = \max_{\substack{\beta_{ij} \in \{0,1\} \\ \xi_i \in \{0,1\}}} \sum_{i=1}^{n} \xi_i$$
$$\text{s.t.} \quad \begin{cases} \sum_{j \neq i} \beta_{ij} \leq (1 - \xi_i) & \forall i = 1, \ldots, n \\ 2 \sum_{j=1}^{n} w_{ji} \beta_{ji} \geq \xi_i \left(\sum_{j=1}^{n} w_{ji} + \rho \right) & \forall i = 1, \ldots, n, \end{cases} \tag{20}$$

where the binary variable ξ_i determines whether the corresponding second constraint is restrictive or trivially satisfied. A suboptimal solution to this problem is still a lowerbound. This motivates the following approximative methodology: In the first step, a convex linear programming problem is solved:

$$(\hat{\beta}_{ij}, \hat{\xi}_i) = \arg\max_{\beta_{ij}, \xi_i} \sum_{i=1}^{n} \xi_i \quad \text{s.t.} \quad \begin{cases} \sum_{j \neq i} \beta_{ij} \leq (1 - \xi_i) & \forall i = 1, \ldots, n \\ 2 \sum_{j=1}^{n} w_{ji} \beta_{ji} \geq \xi_i \left(\sum_{j=1}^{n} w_{ji} + \rho \right) & \forall i = 1, \ldots, n \\ \xi_i \in [0,1] & \forall i = 1, \ldots, n \\ \beta_{ij} \in [0,1] & \forall i, j = 1, \ldots, n. \end{cases} \tag{21}$$

In the second stage, the estimates $\hat{\beta}$ and $\hat{\xi}$ are rounded to the nearest integer solutions satisfying $\beta_{ij} \in \{0,1\}$ and $\xi_i \in \{0,1\}$ gives a suboptimal solution. The cost resulting from this suboptimal procedure associates to the rounded arguments (say $\tilde{\theta}(\rho)$) is guaranteed to be lower or equal to $\theta(\rho)$ by construction.

This reasoning is summarized in the following proposition:

Proposition 2 (Lower-bound to K-Capacity)

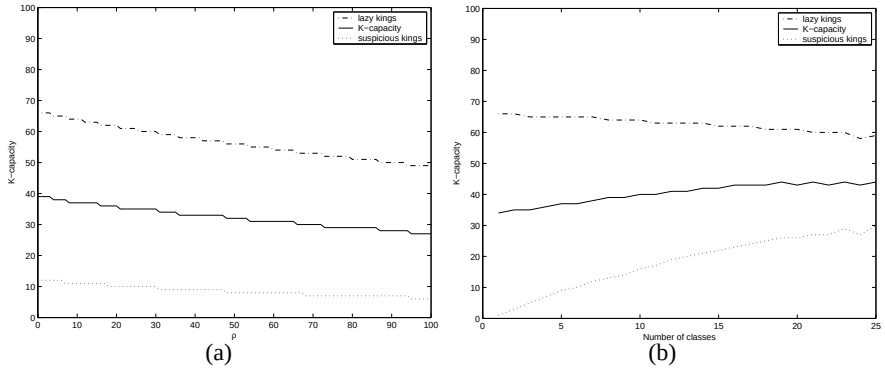
$$\vartheta(\rho) \ \geq \ \theta(\rho) \ \geq \ \tilde{\theta}(\rho). \tag{22}$$

4 ARTIFICIAL EXAMPLE

We discuss the merit of the K-capacity measure, based on a numerical case study. Consider a random graph consisting of 100 nodes, where the first fifty have a label $+1$ and the last 50 label -1. The weights are generated by the following mechanism:

$$w_{ij} \sim \begin{cases} B(p) & \text{iff } \ C(v_i) = C(v_j) \\ B(1-p) & \text{otherwise,} \end{cases} \tag{23}$$

where $B(p)$ denotes a Bernoulli distribution with parameter $0.5 < p < 1$, and $C(v_i)$ denotes the class label of node v_i. This generating mechanism is motivated as follows: nodes from the same class are likely to be strongly connected (within), while connections between the two classes are depreciated. Remark that this is another mechanism as the one discussed



(a) (b)

Fig. 1. *Results of an artificial example.* **(a)** *Display of the K-dimension, the lower- and upper-bound in function of the margin ρ for a random graph with 100 nodes.* **(b)** *Display of the K-dimension, the lower- and upper-bound in function of the number of different classes in the generating mechanism. It is seen that increasing both the margin ρ as well as the number of underlying different classes will make the lower- and upperbound more tight.*

in the transductive setting in Subsection 2.1. The second example employs the same mechanism, but generalizes to a range of different classes.

Figure 1.a gives the result of a simulation study where one graph was generated as indicated, for various parameters ρ ranging from 0 to 50, and for 100 nodes. Here we consider the case of two different classes, a parameter $p = 0.75$. The figure displays the K-capacity as a function of the parameter ρ. The actual capacity was computed by naive (and time-consuming) enumeration, while the upper- and lowerbound respectiovely follow from the linear programming formulation as discussed in the previous section. Figure 1.b gives the result of a study of a set of random graph with 100 nodes, and a ranging number of classes rangin from 1 to 25. Again, both the actual K-capacity as the lower- and upper-bound discussed above are displayed as a function of the number of underlying classes.

5 DISCUSSION

We discussed a measure of capacity of a graph which characterizes the range of labelings which are self-consistent with a certain margin. Furthermore, this paper indicated how this capacity measure can be used to give probabilistic guarantees for learning in a transductive setting. In general, we described a relationship between a probabilistic approach of learning on the one hand, and combinatorial tasks such as graph cut on the other. Specifically, we indicated how the crucial concept of capacity control and regularization in learning can be mapped to graph labeling and graph cuts by using a proper extension operator.

A most interesting question implied by this setup is whether some statements can be made of learnability when $n \to \infty$. Another intruiging matter is whether this capacity concept can be used to characterize the behavior of real world networks as e.g. based on the internet or on citation databases for scientific literature. We hope to get some insight in whether a typical graph grows *inwards* (filling the missing connections) or *outwards* (exploring new regions). From the theoretical point of view, this short paper studied a number

of usefull concepts (including a formal notion of risk, a hypothesis set and an extension operator) which outline an approach towards the study of induction in the semisupervised setting.

References

1. M. Anthony and P.L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
2. A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 19–26. Morgan Kaufmann Publishers, 2001.
3. A. Blum, J. Lafferty, M.R. Rwebangaria, and R. Reddy. Semi-supervised learning using randomized mincuts. *24e International Conference on Machine Learning (ICML)*, 2004.
4. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
5. L. Lovasz. On the Shannon capacity of a graph. *IEEE Transactions on Information Theory*, 25:1–7, 1979.
6. R. El-Yaniv P. Derbeko and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.
7. R.J. Serfling. *Approximation Theorems of Mathematical Statistics*. John Wiley & Sons, 1980.
8. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE transactions on Pattern Recognition and Machine Intelligence*, 22(8), aug. 2000.
9. V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.