Information Measures of Frequency Distributions with an Application to Labeled Graphs

Cliff Joslyn and Emilie Purvine

Abstract The problem of describing the distribution of labels over a set of objects is common in many domains. Cyber security, social media, and protein interactions all care about the manner in which labels are distributed among different objects. In this paper we present three interacting statistical measures on label distributions, thought of as integer partitions, inspired by entropy and information theory. Of central concern to us is how the open- versus closed-world semantics of one's problem leads to different ways that information about the support of a distribution is accounted for. In particular, we can consider the number of labels seen in a particular data set in relation to both the number of items and the number of labels available, if known. This will lead us to consider both two alternate entropy normalizations, and a new measure specifically of support size, based not on entropy but on nonspecificity measures as used in nontraditional information theory. The entropy- and nonspecificity-based measures are related in their ability to index integer partitions within Young's lattice. Labeled graphs are discussed as a specific case of labels distributed over a set of edges. We describe a use case in cyber security using a labeled directed multigraph of IPFLOW. Finally, we show how these measures respond when labels are updated in certain ways corresponding to particular changes of the Young's diagram of an integer partition.

Keywords Information measures · Distributions · Entropy · Nonspecificity · Labeled graph

Mathematics Subject Classification 94A17 · 05C90 · 05A17

C. Joslyn · E. Purvine (✉)
Pacific Northwest National Laboratory, Seattle, WA 99352, USA
e-mail: Emilie.Purvine@pnnl.gov

C. Joslyn
e-mail: Cliff.Joslyn@pnnl.gov

G. Letzter et al. (eds.), *Advances in the Mathematical Sciences*, Association for Women in Mathematics Series 6, DOI 10.1007/978-3-319-34139-2_19

1 Introduction

Given a nonempty collection of labeled entities, how are we to measure the distribution of their labels as drawn from a set of available discrete attributes? Of course, such a basic question should, and does, have many answers already. In particular, entropy measures are commonly used to measure the spread and shape of a distribution, but are defined only on probability distributions. And while probability distributions, as relative frequencies, are easily and uniquely derived from counts, nonetheless this transformation loses information about the support relative to the original count distribution.

We have sought in the literature measures which completely characterize the size and shape of the distribution of labels of such a collection, but were surprised not to find them. We present candidates here, incorporating two forms of uncertainty-based information measures. Entropy and normalized entropy are retained as standard probabilistic approaches over a fixed support, which we introduce as a measure of the "smoothness" of a count distribution. But we also introduce a "dispersion" measure of the degree of support itself relative to the total count. This is a kind of normalized nonspecificity measure, a non-probabilistic uncertainty-based information measure.

In Sect. 2 we introduce integer partitions, different ways to represent them, and our inspiration from collections of labeled items. Additionally, in this section we introduce entropy measures on discrete probability distributions. Next, in Sect. 3 we introduce our two candidate functions measuring the *smoothness* and *dispersion* of a distribution of labels. Then in Sect. 4 we narrow to the case of integer partitions derived from labeled degree distribution where each part counts the number of edges with a particular label. We additionally give a use case related to cyber security. Finally, in Sect. 5 we explore the theoretical aspects of our two functions as they relate to integer partitions and Young diagrams.

2 Preliminaries

In our Preliminaries section we introduce three separate concepts: frequency distributions, integer partitions, and information measures. These three somewhat disjoint concepts will be more strongly related in Sect. 3 and beyond.

2.1 Frequency Distributions

We begin by considering a finite set $X = \{x_j\}_{j=1}^{n}$ of n items, each with a label $l \in \Lambda$, drawn from a set of $\mu = |\Lambda|$ labels. Our goal in this paper is to study the distribution of these μ labels over the set X. We begin then, by forming a frequency vector $\gamma = \langle \gamma_k \rangle_{k=1}^{\mu}$, where $\gamma_k \in \mathbb{N}$ is the number of items in X that are labeled l_k. Notice

here that although we have μ labels available to us we may not use all of the labels, so there might be some $\gamma_k = 0$ in our frequency distribution. In fact, it may be the case that $\mu > n$ so that there *must* be zeros in our distribution as it's impossible to use more than n labels on a set of n objects.

2.2 Integer Partitions

An **integer partition**, or simply a **partition**, of a positive integer n is a list of strictly positive integers, $\lambda = \langle \lambda_1, \lambda_2, \ldots, \lambda_m \rangle$ such that $\lambda_i \geq \lambda_{i+1}$ and $\sum_i \lambda_i = n$. Each λ_i is called a **part**, and we denote the number of parts in λ by $m := |\lambda|$. For example, $\lambda = \langle 5, 5, 3, 1 \rangle$ is a partition of 14 with $m = 4$ parts.

There are many areas in which integer partitions may arise, but we are concerned with those induced by a collection of labeled items. Consider again a collection $X = \{x_j\}_{j=1}^{n}$ of n items, each with a label drawn from a set of labels Λ. When we introduced frequency distributions in the previous section we did our counting "label-wise". Here, to form an integer partition, we will count "item-wise". The labels l_i naturally partition X into $m \leq n$ disjoint blocks $X_i \subseteq X$ where for each $1 \leq i \leq m$, all the labels for the items in X_i are the same. Let $\lambda_i := |X_i|$, $1 \leq i \leq m$ be the number of items in block i, so that $1 \leq \lambda_i \leq n$. Then sort them down, so that $\lambda_i \geq \lambda_{i+1}$. Finally, let $\lambda := \langle \lambda_i \rangle_{i=1}^{m}$. λ is now an integer partition representing the counts of the m labels actually seen, amongst the n items in question.

The item-wise label frequency distribution λ created from X and Λ essentially counts the same thing as the label-wise distribution γ, so let's consider the relationship of γ to λ. Clearly $\mu \geq m$, so $\lambda \subseteq \gamma$ in the sense that each $\lambda_i \in \lambda$ has a corresponding $\gamma_k \in \gamma$, but γ must have $\mu - m$ additional $\gamma_k = 0$. Basically, γ is λ padded with zero counts for any unused labels. Note that if $\mu = m$ then $\gamma = \lambda$ will be a proper integer partition. Therefore, we may refer to γ in the general case below, and λ when we rely on having an integer partition. We will return later to a discussion of γ versus λ in the case of information measures on label distributions, but for now we return to our discussion on integer partitions.

There are various ways of representing partitions either as decreasing lists of numbers as already stated, or as diagrams. A **Young diagram** of a partition λ is an array of boxes which is top and left aligned in which each row represents a single part of the partition. An example is shown in Fig. 1. We will often refer to both the Young diagram of a given partition, and the partition itself, by the same name, e.g., λ. Sometimes these are also called **Ferrers diagrams**, though Ferrers diagrams are often drawn with dots instead of boxes. We will come back to Young diagrams in Sect. 5.

There is a great deal of research in the theory of partitions. Most prominent are enumerative combinatorics questions of the form "How many partitions are there that have a specific property, as a function of n?" Many properties are explored including restricting the number of parts, type of parts (e.g., even parts, distinct parts), and other more complicated types of restrictions.

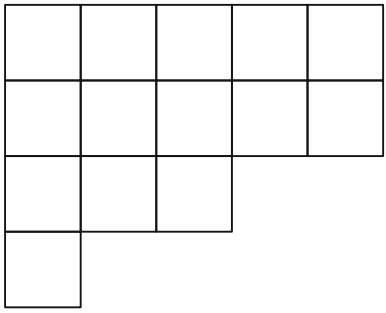

Fig. 1 The Young diagram corresponding to partition $\langle 5, 5, 3, 1\rangle$ of 14

Our research is looking not at these enumerative questions, but instead at calculating information measures, to be introduced in the following sections, over the set of integer partitions, and studying the distribution of these statistics as n, m, and λ_1 (the largest part) vary. Specifically we are interested in characterizing these statistics over the integer partition poset,[1] P_n, referenced in [1, 9]. P_n is the poset of integer partitions of n ordered by refinement. So, if $\eta \leq \lambda$ in P_n then λ has fewer parts than η, and each part in λ is split into multiple parts in η. For example, $\eta = \langle 5, 4, 2, 1, 1, 1\rangle$ is a refinement of $\lambda = \langle 5, 5, 3, 1\rangle$ since one of the 5 s is split into 4, 1, the 3 is split into 2, 1, and everything else remains the same. The elements of P_n can be grouped by their number of parts m, making P_n a *graded* poset [1, 6]. An example of the Hasse diagram of the integer partition poset P_6 is shown in Fig. 2. A Hasse diagram is a visual representation of the cover relations in a poset. In general, each element $p \in P$ from the poset $\mathscr{P} = (P, \leq)$ is shown, and an edge from p to q with p below q indicates that $p \prec q$. We call each group of partitions with the same number of parts a *rank* of P_n. Notice that the ranks of P_n are separated onto different vertical levels of the Hasse diagram.

Finally, note that the elements of each integer partition poset P_n sit within Young's lattice, a lattice of all integer partitions ordered by containment of their Young diagrams. One Young diagram, η, is contained within another, λ, if every box in η occurs in λ. In other words, if we label each box with two integer coordinates, the x coordinate increasing along columns and the y increasing along rows (as in labeling matrix entries), then every label that occurs in η will also appear in λ. The Young's lattice for $n \leq 6$ is shown in Fig. 3, the dashed edges indicate partition refinements in Young's lattice produced by adding or deleting boxes in a Young diagram λ, and thus a change in n. Solid edges indicate partition refinements within each of the P_n produced by splitting or merging blocks, and thus a change in m, but with no change in n, these solid edges are not included in Young's lattice. The significance of the G notations at the bottom of Fig. 3 will be discussed later in Sect. 3. Notice that Young's lattice is also graded, with each rank being the set of partitions in one P_i.

[1] The integer partition poset should not be confused with the *set* partition *lattice* [1].

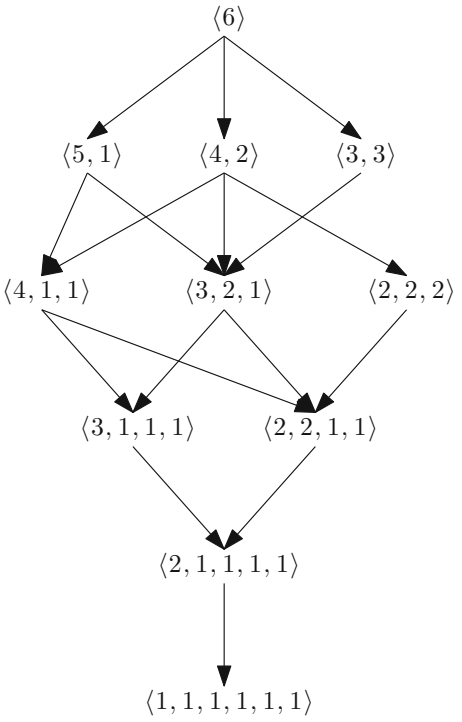

Fig. 2 The Hasse diagram of the integer partition poset for partitions of $n = 6$, P_6

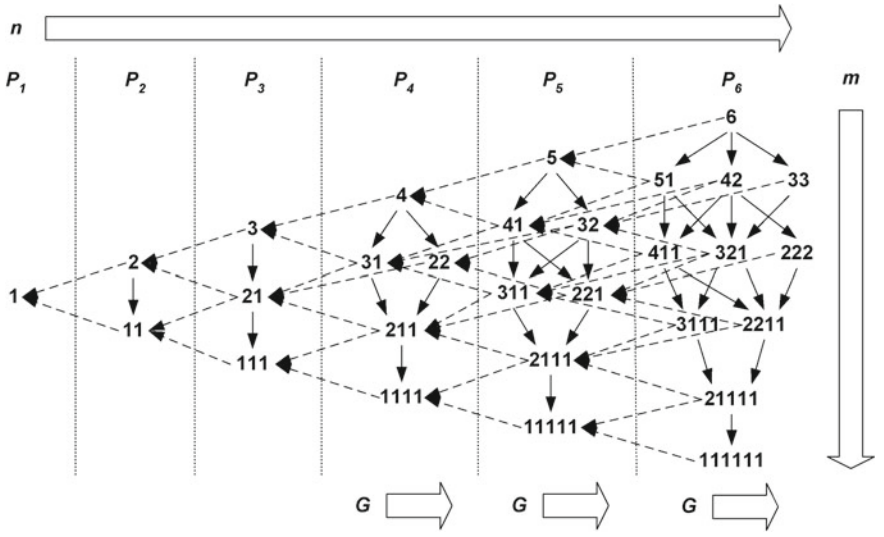

Fig. 3 Partition posets P_i (*solid arrows*) embedded as ranks i within Young's lattice (*dashed arrows*)

2.3 Information Measures

Let $\mathbf{p} = \langle p_1, p_2, \ldots, p_\mu \rangle$ be a discrete probability distribution, where the $p_k := p(y_k) \in [0, 1]$ are the probabilities of a random variable Y taking values in the set $\{y_1, \ldots, y_\mu\}$, so that $\sum_{k=1}^{\mu} p_k = 1$. The **entropy** of the distribution $\mathbf{p}$, denoted $H(\mathbf{p})$, is given by the following equation:

$$H(\mathbf{p}) = -\sum_{k=1}^{\mu} p_k \log_2 p_k.$$

In the case where any $p_k = 0$ we define $0 \log_2(0) := 0$. This is a logical definition since $\lim_{x \to 0} x \log_2(x) = 0$. There are continuous analogs of entropy where the sum is replaced by an integral, but we only deal with the discrete version in this paper.

Entropy can be thought of as a measure of uncertainty in a probability distribution. If the distribution is uniform, $\mathbf{p}_u = \left\langle \frac{1}{\mu}, \ldots, \frac{1}{\mu} \right\rangle$, then the entropy is $H(\mathbf{p}_u) = -\log_2(\frac{1}{\mu}) = \log_2(\mu)$. This represents maximum uncertainty, i.e., any outcome is equally likely. Contrast that with the fully skewed distribution $\mathbf{p}_s = \langle 0, \ldots, 0, 1 \rangle$ which has entropy $H(\mathbf{p}_s) = -\log_2(1) = 0$. This represents no uncertainty since the outcome is determined with all of the probability mass sitting on only one possibility.

Normalizing entropy by its maximum, $\log_2(\mu)$, is a standard approach to effectively measure the shape of the probability distribution. We call this normalized entropy **p-smoothness** of a probability distribution $\mathbf{p}$ and denote it by $\widetilde{G}(\mathbf{p})$:

$$\widetilde{G}(\mathbf{p}) = \frac{H(\mathbf{p})}{\log_2(\mu)} = \frac{-\sum_{k=1}^{\mu} p_k \log_2 p_k}{\log_2(\mu)} \in [0, 1]. \tag{1}$$

Just as in the definition for entropy, we must define $\widetilde{G}(\mathbf{p})$ in a special case. This time we treat $\mu = 1$, which would require dividing by $\log_2(1) = 0$. In this case we let $\widetilde{G}(\langle 1 \rangle) := 1$ which agrees with $\lim_{x \to 1} \frac{x \log_2(x)}{\log_2(x)} = 1$. We are using the word "smooth" here as a synonym for "close to uniform". We know that a uniform distribution has maximum entropy and therefore, when normalized, $\widetilde{G}(\mathbf{p}) = 1$. So, a highly smooth distribution is one that is close to uniform. On the other hand, a very skewed distribution, one that looks "lumpy", will have low entropy (low uncertainty) and therefore $\widetilde{G}(\mathbf{p})$ smaller. So, the closer the distribution is to perfectly smooth, or uniform, the higher $\widetilde{G}(\mathbf{p})$ will be.

3 Information Measures on Label Distributions

In this paper we are concerned with measuring the shape of a distribution of labels over a set of items. Specifically, given an arbitrary collection of n labeled items $X = \{x_j\}_{j=1}^{n}$, how can we best characterize the distribution of their labels Λ? Now,

this distribution is not a probability distribution but instead it is an *absolute* frequency distribution. Classically, the first thing done is to transform the associated label-wise distribution, γ, described in Sect. 2.2, which we recall may be a proper integer partition, λ, into a *relative* frequency distribution $\mathbf{f}$ as

$$\mathbf{f}(\gamma) := \gamma/n = \langle f_k \rangle_{k=1}^{\mu} = \langle \gamma_k/n \rangle_{k=1}^{\mu}$$

so that now $f_k \in [0, 1]$ and $\sum_{k=1}^{\mu} f_k = 1$. Relative frequency distributions $\mathbf{f}$ are typically interpreted as (discrete) probability distributions $\mathbf{p}$ by considering f_k as the probability of choosing an item labeled l_k when picking an item randomly from the set of n labeled items, $X = \{x_j\}$. Therefore, we can calculate the entropy and $\mathbf{p}$-smoothness as introduced in the previous section. With slight abuse of notation here we define

$$\widetilde{G}(\gamma) := \widetilde{G}\left(\mathbf{f}(\gamma)\right) = \frac{H(\mathbf{f}(\gamma))}{\log_2(\mu)} = \frac{-\sum_{k=1}^{\mu} \frac{\gamma_k}{n} \log_2 \frac{\gamma_k}{n}}{\log_2(\mu)}.$$

Now, consider our set X with items labeled from Λ and create both the integer partition λ item-wise and the distribution γ label-wise. We can treat λ similarly, taking

$$\mathbf{f}(\lambda) := \lambda/n = \langle f_i \rangle_{i=1}^{m} = \langle \lambda_i/n \rangle_{i=1}^{m}.$$

$\mathbf{f}(\lambda)$ is basically $\mathbf{f}(\gamma)$ stripped of its $\mu - m$ trailing zeros. And then we have

$$\widetilde{G}(\lambda) := \widetilde{G}\left(\mathbf{f}(\lambda)\right) = \frac{H(\mathbf{f}(\lambda))}{\log_2(m)} = \frac{-\sum_{i=1}^{m} \frac{\lambda_i}{n} \log_2 \frac{\lambda_i}{n}}{\log_2(m)}.$$

We observe that if $m < \mu$ then $\widetilde{G}(\gamma) \neq \widetilde{G}(\lambda)$ despite being calculated from the same sets X and Λ. Non-normalized entropy H, the numerator of $\widetilde{G}$, cannot distinguish the relative frequency distributions of λ and γ as it is blind to zero-padding, but the normalization factor is different in each case, $\log_2(\mu)$ on the left-hand side and $\log_2(m)$ on the right.

The question of whether or not these measures *should* be equal is not for this paper to decide. However, we present an additional measure which is blind to zero-padding on absolute frequency distributions and later will show that there can be advantages in using one or the other. We define **smoothness** as

$$G(\gamma) = G\left(\mathbf{f}(\gamma)\right) := \frac{H\left(\mathbf{f}(\gamma)\right)}{\log_2(m)} = \frac{-\sum_{i=1}^{\mu} \frac{\gamma_k}{n} \log_2 \frac{\gamma_k}{n}}{\log_2(m)}.$$

Notice the difference being that here we normalize by the number of nonzero entries in γ, and now $G(\gamma) = G(\lambda) = \widetilde{G}(\lambda)$. So, G is only sensitive to the *absolute* support, m, of λ or γ, that is, the number of labels actually *seen*, as opposed to the number μ of labels which are *available*. Another way of saying this is that $\widetilde{G}$, as opposed to G,

makes a *closed-world* assumption that we know in advance the universe of discourse Λ of available labels. In open-world situations, Λ may be unknown, unspecified, or so large as to be meaningless to the problem being modeled. The open world assumption of G is thus independent of any implicit assumptions about the space of labels.

While the open-world smoothness G provides an important alternative to the traditional closed-world normalized entropy $\widetilde{G}$, still for our question of characterizing counts of labels in the context of integer partitions λ, or more general label distributions γ, both G and $\widetilde{G}$ have flaws. Both relative frequency distributions $\mathbf{f}(\gamma), \mathbf{f}(\lambda)$ lose information relative to absolute frequency distributions γ, λ. In particular, consider two label frequency distributions γ and γ' of n and n' objects. Let $\gamma' = \alpha \cdot \gamma$ for some $\alpha \in \mathbb{N}$, so that, incidentally, $n' = \alpha n$. Then additionally $\mathbf{f}(\gamma) = \mathbf{f}(\gamma')$, and of course, both $G(\gamma) = G(\gamma')$ and $\widetilde{G}(\gamma) = \widetilde{G}(\gamma')$. In reducing to relative frequencies, we have lost connection to the interpretation of the partitions as counts with respect to a total number of items n, n', and thus neither smoothness G nor $\mathbf{p}$-smoothness $\widetilde{G}$ can distinguish these cases.

We have observed now that both G and $\widetilde{G}$ are insensitive to the total number of items that we are labeling, $n = |X|$, and additionally that $\widetilde{G}$ distinguishes between the distributions item-wise (λ an integer partition) and label-wise (γ an arbitrary non-negative integer distribution) when we have more labels available, μ, than observed, m, whereas G does not. It is not our goal in this paper to decide which of $\widetilde{G}$ and G is more reasonable to use, indeed it often depends on the application. In some cases, and often in those that we are concerned with, e.g., cyber networks, either the number of possible labels is very large compared to the number of labels seen, or we simply do not know how many labels are possible. In these cases of large μ normalizing with respect to μ gives little to no information as all $\widetilde{G}$ values will then be very small. It is for these reasons that we generally turn our attention in this paper to the use of G over $\widetilde{G}$, or when we do use $\widetilde{G}$ we will often make the assumption that $\mu = n$. We note that we will contrast G and $\widetilde{G}$ again later using this assumption on μ when we introduce a third measure called κ. Additionally since $G(\lambda) = G(\gamma)$ we will proceed working with only integer partitions.

As evidenced by the prior discussion on G versus $\widetilde{G}$ it's clear that there is an important relationship between the three quantities n, m, and μ. While the number of items n and the number of available labels μ themselves need not be related, we do have $m \leq \min(n, \mu)$; that is, there can be as many labels, m, in X as the number of items n, unless $\mu < n$, in which case there are not enough labels to go around, and some items x_i must "double up." So, there are two edge cases:

$m = n$: Here all items have a distinct label, so that $\lambda = \langle 1, 1, \ldots, 1 \rangle$. Also $n \leq \mu$, and our label-wise frequency distribution is $\gamma = \langle 1, 1, \ldots, 1, 0, 0, \ldots 0 \rangle$, with m ones and $\mu - m$ zeroes.

$m = \mu$: Here all available labels are used, so that $n \geq \mu$, and $\lambda = \gamma$ with all the $\lambda_k \geq 1$.

Consider P_6 as shown in Fig. 4. The integer partitions $\lambda \in P_6$ are shown adorned with the values for G, and m is shown on the right. The κ values on the right of this

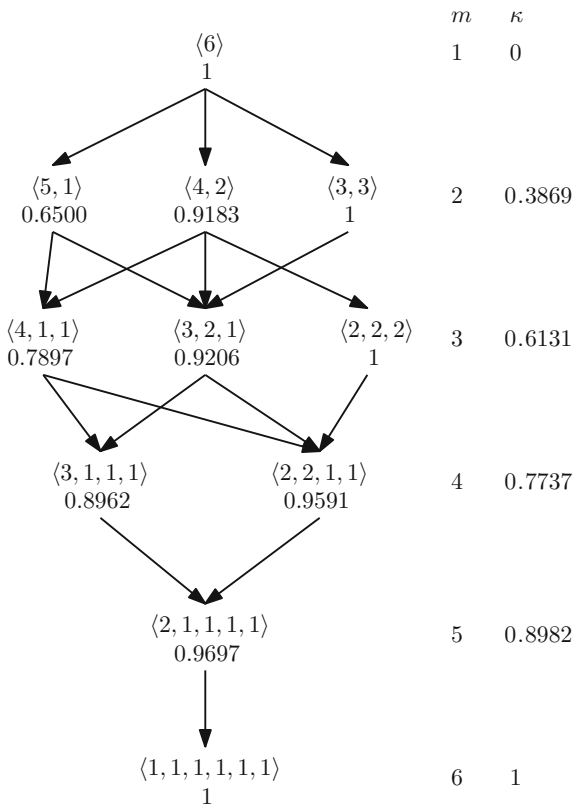

Fig. 4 The integer partition P_6 adorned with smoothness $G(\lambda)$

figure will be defined later. As a general matter, for a given n, each level of P_n includes partitions with the same number of parts. $G(\lambda)$ then orders the λ within each level, with $G(\lambda) = 1$ iff λ is uniform, and $G(\lambda) \to 0$ for $\lambda = \langle n - k, 1, \ldots, 1\rangle$ for each $1 \leq k \leq n - 1$ as $n \to \infty$. Note that in some cases we have $G(\lambda_1) = G(\lambda_2)$ when $\lambda_1 \neq \lambda_2$, e.g. when $n = 20$ and $m = 6$ we have $G([8, 3, 3, 2, 2, 2]) = G([6, 4, 4, 4, 1, 1])$. We will order these partitions according to lexicographic ordering when they occur.

The role of G to order partitions is illustrated in Fig. 3 in Young's lattice, with G operating within each P_n. So more generally, for any partition λ of any n, we seek measures to place it within Young's lattice in terms of:

- Its rank within Young's lattice, which is clearly just n.
- Within each rank of Young's lattice (that is, within P_n), its "horizontal" placement. This is its smoothness $G(\lambda)$.
- And again within each P_n, its "vertical" level within P_n.

So concerning this final quantity, we are left with the question, is there a way to order the integer partitions based on how many parts they have? The number of parts, m,

is the obvious answer, as together with G this ordering would essentially create a coordinate system on the integer partition poset P_n.

We could simply measure the number of parts m relative to the number of possible parts n, whether or not some absolute number of labels μ set an upper bound on m. For a partition of n with m parts this would simply seem to be $\frac{m}{n}$. However, the distribution of the number of partitions of n with a given number of parts has a long tail, i.e., there are relatively few partitions of n with large numbers of parts, and we seek a measure which has larger gaps when there are many partitions, when m is smaller, and smaller gaps when there are fewer partitions, rather than something linear in the number of parts. Additionally, while working with information measures like G, there are great advantages of working with log functions, due to additivity and other properties. We therefore use a log ratio and define what we call **dispersion**, denoted by $\kappa(\lambda)$, as

$$\kappa(\lambda) = \frac{\log_2(m)}{\log_2(n)} \in [0, 1].$$

Figure 4 also shows the κ value for each level of P_6. Like G, we also have $\kappa(\lambda) \in [0, 1]$, but now $\kappa = 0$ if and only if $\lambda = \langle n \rangle$ which is the case only when $m = 1$. Additionally, $\kappa = 1$ if and only if $\lambda = \langle 1, 1, \ldots, 1 \rangle$ when $m = n$. Notice that $\kappa = 1$ implies $G = 1$, but we can have $G = 1$ with small κ, as in $\lambda = \langle 3, 3 \rangle \in P_6$.

Note that unlike G, the value of κ does *not* depend on the actual *values* of the $\lambda_i \in \lambda$, but only on m, the *number* of parts into which λ is divided, and n the total sum of λ. It is effectively a measure of the *support* of the partition λ of the integer n *relative* to n itself. And while G, as an entropy, is a measure of the information content of the partition λ, interpreted as a relative frequency (that is, discrete probability) distribution, $\mathbf{f}(\lambda)$, so κ is *also* a measure of information, although not an entropy, but rather a **Hartley** measure [5].

In the context of **generalized information theory**, a Hartley measure is an information measure called a **nonspecificity N**. Given a collection of m choices, then their nonspecificity is simply $\mathbf{N}(m) = \log_2(m)$. Note that this quantity $\log_2(m)$ is also the maximal entropy $H(\mathbf{p})$ when $\mathbf{p}$ is uniform, and in this particular case these measures coincide. But like entropy, in more general cases nonspecificities can take on more complex forms, and are fully-fledged information measures in that they satisfy basic axioms of additivity, monotonicity, and normalization as they quantify the amount of uncertainty present in a collection of choices which are not probabilistically weighted.

In our case, our $\kappa = \log_2(m)/\log_2(n)$ is thereby a **normalized nonspecificity**. As a general matter, nonspecificities are defined and used in the context of possibilistic information theory, or possibility theory [2, 3, 8]. Although exploring possibilistic measures of support is not the purpose of this paper, here it is sufficient to observe that as entropy measures the probabilistic constraint placed on a collection of m choices by the probabilities f_i, so κ measures the non-probabilistic constraint placed on a collection of n choices by the selection of m of them.

We have already observed that we can use κ and G as two "coordinates" within the Hasse diagram of the integer partition poset. We use κ to tell the vertical level of a partition, a y coordinate in the Hasse diagram, and G as an x coordinate telling how far to the right (closer to uniform) a partition is. This is illustrated in Fig. 5(a) where κ (y-axis) and G (x-axis) are calculated on all partitions of $n = 20$. We see clear levels corresponding to each m value yielding different κ values. The G values then go up to a maximum of nearly 1 depending on if m divides 20 or not.

Another important observation is as follows. In our "typical case" of $n = \mu$ so that we have no more labels than there are items, we actually have that

$$\widetilde{G} = G \cdot \kappa.$$

This, together with $\kappa \in [0, 1]$, shows that $\widetilde{G}$ can be interpreted as both a weighting down of G by κ, and simply as a quantity reflecting both smoothness and dispersion through the product, capturing information from both G and κ. So if $\widetilde{G}$ captures information from both κ and G, the question is: can we use it to determine both the x and y coordinates of a partition? In order for this to be the case we would need to be able to break up the range of $\widetilde{G}$, [0, 1], into disjoint intervals, one for each level of the poset (m value). Unfortunately, we can only do this for $n \leq 7$. In the integer partition poset for $n = 8$ we cannot decompose the interval [0, 1] into disjoint intervals for each m value. For $m = 3$ the $\widetilde{G}$ values range between 0.3538 and 0.5204, and for $m = 4$ the values range between 0.5163 and 0.6667. There is a clear overlap between the two ranges, so we cannot in general use $\widetilde{G}$ to decide which level a partition is on. The plot in Fig. 5b shows the plot of κ now against $\widetilde{G}$ (with $n = \mu$) to show the overlap in $\widetilde{G}$ ranges as it is broken up into its κ levels.

Considering the relation between our three measures G, $\tilde{G}$, and κ, we can conclude that it is more enlightening to use both κ and G separately to give information about

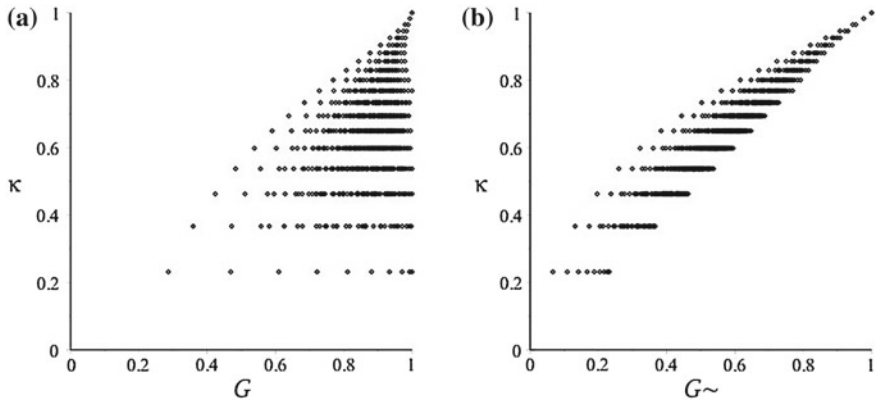

Fig. 5 Plots showing how G, κ, and $\widetilde{G}$ interact for all partitions of $n = 20$. **a** κ versus G. **b** κ versus $\widetilde{G}$ with $\mu = n$

a partition λ. But at the same time, it might be convenient to use $\widetilde{G}$ as an alternate entropy normalization, a discounting of G by κ, or a scalar combining G and κ. Effectively, $\widetilde{G}$ provides a single, scalar quantity reflecting both the horizontal and vertical position within P_n. Together with n itself, we can use them to uniquely characterize any integer partition λ within Young's lattice.

4 Application to Labeled Degree Distributions

So far in this paper we have described dispersion (κ) and smoothness (G) as measures on integer partitions, λ, or more specifically on frequency distributions of a set of labels, Λ, on a set of objects X. Now we come to the application of labeled graphs. Consider a directed graph or multigraph, $G = (V, E)$, with edge label function $f : E \to \Lambda$. Theoretically, any collection of edges can be considered as our set of objects, X, but we will restrict to the case where the set of edges has a common source vertex or target vertex. Given a vertex, $v \in V$, let $S_v = \{e = e_1 e_2 \in E : e_1 = v\}$ be those edges which have v as their source vertex, and $T_v = \{e = e_1 e_2 \in E : e_2 = v\}$ be those which have v as their target vertex. We may treat S_v and T_v separately as base sets, or take their union and consider the full set of edges incident on v. For example, see Fig. 6.

Given these sets of edges, S_v and T_v, we consider not just the size of the sets, as traditionally considered when investigating degree distribution, but also their dispersion and smoothness with respect to the labeling function f and label set Λ. Referring back to Fig. 6 the integer partition corresponding to T_v is $\lambda_{T_v} := \langle 2, 1, 1\rangle$ where the labels are in the order C, A, D, and integer partition for S_v is $\lambda_{S_v} := \langle 2, 1, 1, 1\rangle$ for labels B, A, C, E. If we wish to take the full set of edges into account we have $\lambda_v := \langle 3, 2, 2, 1, 1\rangle$ for labels C, A, B, D, E. Figure 7 shows the Young diagrams for these three example partitions. Then we can calculate the dispersion, smoothness, and λ-smoothness of these distributions. These values are found in Table 1. As we might expect, λ_{S_v} has the highest smoothness among the three partitions and also the highest dispersion. But, these three partitions are not very diverse and so we get very similar G, $\widetilde{G}$, and κ values.

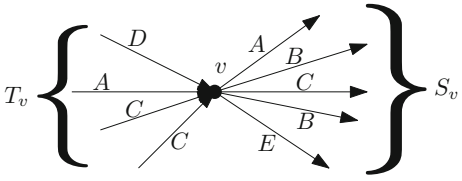

Fig. 6 An example vertex in a labeled directed graph with 4 in-edges and 5 out-edges

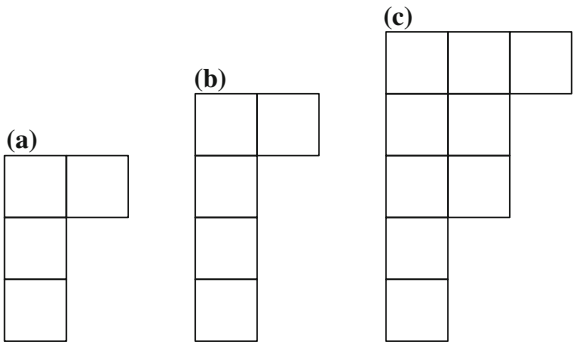

Fig. 7 Young diagrams for our three example partitions. **a** λ_{T_ν}. **b** λ_{S_ν}. **c** λ_ν

Table 1 Smoothness, λ-smoothness, and dispersion values for three example partitions

	λ_{T_ν}	λ_{S_ν}	λ_ν
G	0.9464	0.9610	0.9463
$\widetilde{G}$	0.7500	0.8277	0.6931
κ	0.7925	0.8614	0.7325

4.1 Cyber Security Use Case

This question of assessing the shape of labeled degree distributions, or more generally sets of labeled objects or integer partitions, through information measures was originally motivated by a problem in cyber security. Securing cyber systems has become more and more necessary since attacks to large companies and governments have become increasingly common. Detecting different types of attacks while maintaining system resiliency, i.e., being able to complete missions in the face of attack, is a major focus in cyber security.

We focus on the use of a specific type of cyber data called NetFlow, or more generally IPFLOW, which is IP communication traffic collected at routers and switches throughout the network. A single IPFLOW record contains a source IP and port, a destination IP and port, as well as other data about the information being sent including start and end time, number of packets, and number of bytes. We can study IPFLOW data by transforming it into an IPFLOW multigraph in which vertices are IP addresses, or IP:port pairs, and edges indicate a flow of information from one IP to another. Figure 8 shows an example IPFLOW graph where IPs have been reduced down to their last two octets (e.g., a.x instead of $\alpha.\beta$.a.x), and edges are labeled with a flow ID.

Many common attack types have signatures for the way that they are carried out that can be seen in the IPFLOW graph. Our observation was that these signatures can manifest as extreme shifts in smoothness and dispersion values. For example, consider a denial of service (DoS) or distributed denial of service (DDoS) attack. This occurs when an adversary, or a distributed group of adversaries, floods a server with external communication requests. This overloads the server so that it cannot respond

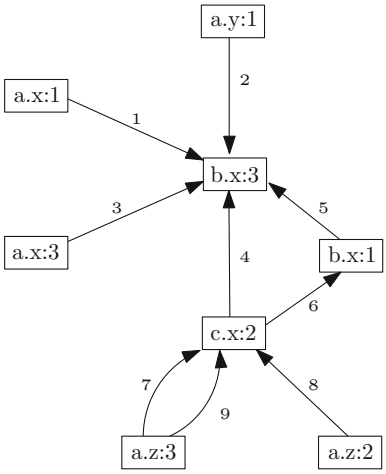

Fig. 8 An example IPFLOW graph with vertices being IP:port pairs and edges labeled with a flow ID

to all of the requests in a timely manner, effectively making the server unavailable to legitimate requests.

If we are looking at a single attacker, IP_A, a DoS attack manifests as a large out-degree with the majority of edges having the same destination. Consider the set of edges with source IP_A and label each edge with its destination IP address. The distribution of these labels will have a single, or very few, labels with high count and any others with very low count. The smoothness of this distribution will likely depend on the number of victims and non-victims being contacted by IP_A. If IP_A only contacts victims then smoothness is likely to be high, but if they are also contacting others (e.g., fellow attackers or a controller in a DDoS attack) then smoothness will be low. However, we can say that dispersion will be very low since the number of communications (the number of edges or items) will be much larger than the number of IPs that are contacted (the number of labels).

In Fig. 9 we show the outgoing smoothness versus outgoing dispersion for vertices in an IPFLOW graph when edges are labeled as described above, by the destination IP address. We used NetFlow from the 2013 Visual Analytics Science and Technology (VAST) Challenge data set which contains synthetic NetFlow with well-documented ground truth attacks [7]. The blue data points are external IP addresses and those large blue points on the left that are labeled are known to be attackers. The cluster of blue points on the upper right, with the IP 10.0.0.5 singled out, are virtual websites. Their high smoothness and dispersion mean that they send information fairly uniformly to other IP addresses, and they do not send very much to each, which is expected behavior for a website. The size of each circle indicates how many flows were sent. We have more results on this data set using these dispersion and smoothness measures as well as other analysis in [4].

Other types of attacks have similar characterizations through smoothness and dispersion. Port scans, for example, where one or many external IPs contact a single vertex through many ports, can be seen again in the IPFLOW graph. The port scanners

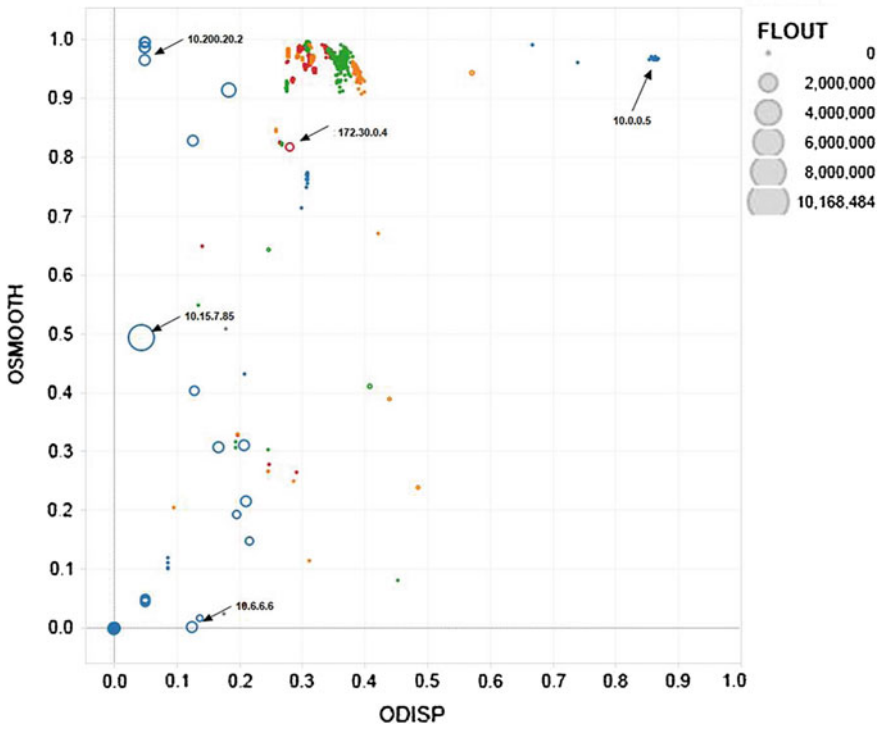

Fig. 9 Outgoing smoothness (y-axis) versus outgoing dispersion (x-axis) for vertices in an IPFLOW graph created by synthetic IPFLOW data around the time of a DDoS attack

will have both high outgoing dispersion and high outgoing smoothness when edges are labeled with destination IP and port. Each scanner contacts each port on a given IP address a small number of times, and does so fairly uniformly.

Though we have only worked in the cyber security use case we can see the value of using these measures on vertices in other domains to enrich the degree distribution. Labeled graphs have a richer set of information than unlabeled graphs and so we should be using that extra information to perform analysis. Other possible application areas could be social network graphs where edge labels are type of communication or type of relationship between people, or protein interaction graphs where distributions come from weights on edges indicating magnitude of interaction.

5 Relation to Young Diagrams

Although we were inspired to create the dispersion and smoothness functions to measure the shape of labeled degree distributions in directed, labeled graphs, we

realize that they are also interesting mathematical functions on the set of integer partitions. In this section we will give a few of our observations about how G and κ change when we change an integer partition in two specific ways.

Recall that a Young diagram is one way to pictorially represent an integer partition as m rows of boxes, the ith row containing λ_i boxes. We were interested in how G and κ vary as we transform the Young diagram. Two transformations will be considered, (1) moving one box from row j to row i, in the case of labeled items this corresponds to changing one label to something else already seen, and (2) taking the conjugate of the diagram.

5.1 Moving Boxes

We will first consider moving one box from row j to row i. Consider two partitions, λ and η, where we form η from λ by switching one box as described. In this case we can write η in terms of λ

$$\eta_\ell = \begin{cases} \lambda_i + 1 & \ell = i \\ \lambda_j - 1 & \ell = j \\ \lambda_\ell & else. \end{cases}$$

We are interested in how G and κ change as we go from λ to η. If we make the stipulation that $\lambda_j \geq 2$ then we are never removing an entire part from λ, therefore $\kappa(\lambda) = \kappa(\eta)$. However, G will change as long as $\lambda_i + 1 \neq \lambda_j$.

$$\begin{aligned}
\log_2(m)\left[G\left(\mathbf{f}(\lambda)\right) - G\left(\mathbf{f}(\eta)\right)\right] &= -\sum_{\ell=1}^{m} \frac{\lambda_\ell}{n} \log_2 \frac{\lambda_\ell}{n} + \sum_{\ell=1}^{m} \frac{\eta_\ell}{n} \log_2 \frac{\eta_\ell}{n} \\
&= -\frac{\lambda_i}{n} \log_2 \frac{\lambda_i}{n} - \frac{\lambda_j}{n} \log_2 \frac{\lambda_j}{n} + \\
&\quad + \frac{\lambda_i + 1}{n} \log_2 \frac{\lambda_i + 1}{n} + \frac{\lambda_j - 1}{n} \log_2 \frac{\lambda_j - 1}{n} \\
&= -\frac{\lambda_i}{n} \log_2 \frac{\lambda_i}{n} + \frac{\lambda_i}{n} \log_2 \frac{\lambda_i + 1}{n} + \frac{1}{n} \log_2 \frac{\lambda_i + 1}{n} + \\
&\quad - \frac{\lambda_j}{n} \log_2 \frac{\lambda_j}{n} + \frac{\lambda_j}{n} \log_2 \frac{\lambda_j - 1}{n} - \frac{1}{n} \log_2 \frac{\lambda_j - 1}{n} \\
&= \frac{\lambda_i}{n} \left[\log_2 \frac{\lambda_i + 1}{\lambda_i}\right] + \frac{\lambda_j}{n} \left[\log_2 \frac{\lambda_j - 1}{\lambda_j}\right] + \\
&\quad + \frac{1}{n} \left[\log_2 \frac{\lambda_i + 1}{\lambda_j - 1}\right]
\end{aligned}$$

We will bound each of these terms independently to see what kind of change in G we can have. First, notice that λ_i cannot be bigger than $n - 1$ since we are adding 1 to it, and λ_j cannot be less than 2 since we are subtracting 1 from it, it also cannot be more

than $n-1$ or there would be nowhere else to move a box. Therefore, $1 \leq \lambda_i \leq n-1$ and $2 \leq \lambda_j \leq n-1$.

Now, let's focus on upper and lower bounds for the first term, $\frac{\lambda_i}{n}\left[\log_2 \frac{\lambda_i+1}{\lambda_i}\right]$. It is not difficult to check that $\frac{d}{dx}\left(\frac{x}{n}\log_2 \frac{x+1}{x}\right)$ is positive for $x \geq 1$, so this term must be increasing. Therefore, we can get upper and lower bounds for the term by plugging in the maximum and minimum values for λ_i, respectively.

$$\frac{1}{n} = \frac{1}{n}\log_2 \frac{1+1}{1} \leq \frac{\lambda_i}{n}\left[\log_2 \frac{\lambda_i+1}{\lambda_i}\right] \leq \frac{n-1}{n}\log_2 \frac{n-1+1}{n-1} = \frac{n-1}{n}\log_2 \frac{n}{n-1}.$$

Next, we look at bounds for the second term, $\frac{\lambda_j}{n}\left[\log_2 \frac{\lambda_j-1}{\lambda_j}\right]$. Again, we can see that the derivative $\frac{d}{dx}\left(\frac{x}{n}\log_2 \frac{x-1}{x}\right)$ is positive for $x \geq 2$ and so we have an increasing function of λ_j. We again plug in maximum and minimum values for λ_j to get upper and lower bounds.

$$-\frac{2}{n} = \frac{2}{n}\log_2 \frac{1}{2} = \frac{2}{n}\log_2 \frac{2-1}{2} \leq \frac{\lambda_j}{n}\left[\log_2 \frac{\lambda_j-1}{\lambda_j}\right] \leq \frac{n-1}{n}\log_2 \frac{n-2}{n} < 0.$$

Finally, we need to get bounds for the third term, $\frac{1}{n}\left[\log_2 \frac{\lambda_i+1}{\lambda_j-1}\right]$. In this case we have a function of both λ_i and λ_j so using the derivative to tell if it is increasing or decreasing will not work. Instead, we bound $\frac{\lambda_i+1}{\lambda_j-1}$ and then take the $\log_2$ of those bounds.

$$\frac{2}{n-2} = \frac{1+1}{n-1-1} \leq \frac{\lambda_i+1}{\lambda_j-1} \leq \frac{(n-1)+1}{2-1} = n.$$

Then, our bounds for $\frac{1}{n}\left[\log_2 \frac{\lambda_i+1}{\lambda_j-1}\right]$ are:

$$\frac{1}{n}\log_2 \frac{2}{n-2} \leq \frac{1}{n}\left[\log_2 \frac{\lambda_i+1}{\lambda_j-1}\right] \leq \frac{1}{n}\log_2 n$$

Putting everything together we can bound our original quantity, $\log_2(m)\left[G\left(\frac{\lambda}{n}\right) - G\left(\frac{\eta}{n}\right)\right]$. The upper bound is given by:

$$\log_2(m)\left[G\left(\frac{\lambda}{n}\right) - G\left(\frac{\eta}{n}\right)\right] \leq \frac{n-1}{n}\log_2 \frac{n}{n-1} + 0 + \frac{1}{n}\log_2 n = \frac{1}{n}\log_2\left(\frac{n^n}{(n-1)^{n-1}}\right).$$

The lower bound is

$$\log_2(m)\left[G\left(\frac{\lambda}{n}\right) - G\left(\frac{\eta}{n}\right)\right] \geq \frac{1}{n} - \frac{2}{n} + \frac{1}{n}\log_2 \frac{2}{n-2} = \frac{1}{n}\log_2 \frac{1}{n-2}$$

Both of these bounds tend to 0 as $n \to \infty$, so that means that in the long run if we just change one bit in the integer partition we don't change G by very much. This makes sense because intuitively as n gets larger, moving one bit, or relabeling one element of X, should impact the smoothness less and less as n gets larger.

5.2 Conjugation

The second way we can transform a Young diagram is by conjugating it. This is flipping it over its main diagonal, as in Fig. 10. The conjugate of λ is written as λ^*. Because κ depends entirely on the number of parts (along with the value n being partitioned), and the number of parts of λ^* is equal to the largest part of λ, we can prove sharp bounds on $\kappa(\lambda^*)$ in terms of $\kappa(\lambda)$.

Proposition 1 *Let λ be an integer partition of n with m parts, and λ^* be its conjugate, with m^* parts. Then*

$$1 - \kappa(\lambda) \leq \kappa(\lambda^*) \leq \frac{\log_2(n - n^{\kappa(\lambda)} + 1)}{\log_2(n)}, \tag{2}$$

and these bounds are sharp.

Proof We have already mentioned the fact that the number of parts of λ^* is equal to the largest part of λ, or λ_1. Therefore, we can write $\kappa(\lambda^*)$ in terms of the largest part of λ as

$$\kappa(\lambda^*) = \frac{\log_2(m^*)}{\log_2(n)} = \frac{\log_2(\lambda_1)}{\log_2(n)}. \tag{3}$$

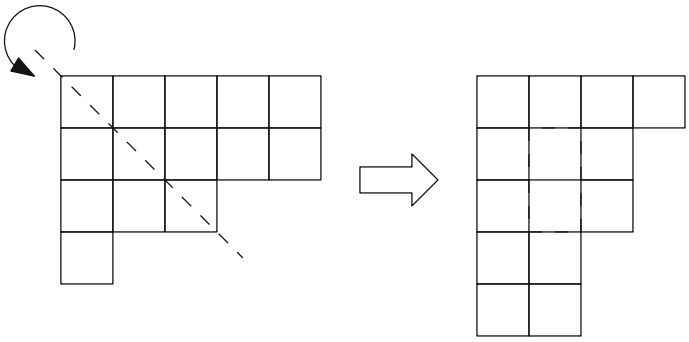

Fig. 10 The conjugate of a partition is created by flipping the associated Young diagram over the dotted diagonal line. For example, the conjugate of $\lambda = \langle 5, 5, 3, 1 \rangle$ is $\lambda^* = \langle 4, 3, 3, 2, 2 \rangle$

Now, observe that we can bound λ_1 in terms of m. We claim an upper bound $\lambda_1 \leq n - m + 1$. If λ_1 were strictly larger than $n - m + 1$ and all other parts were equal to 1 (the smallest they can be) then we would have

$$n = \sum_{i=1}^{m} \lambda_i = \lambda_1 + (m-1) \cdot 1 > (n - m + 1) + m - 1 = n$$

which is a contradiction. This upper bound is sharp which can be seen when $\lambda_1 = n - m + 1$ and all other parts are equal to 1.

Then, we claim a lower bound of $\lambda_1 \geq \frac{n}{m}$. Here if instead $\lambda_1 < \frac{n}{m}$ then we would also have $\lambda_i < \frac{n}{m}$ for all i since the parts are in decreasing order. Then in this case

$$n = \sum_{i=1}^{m} \lambda_i < m \cdot \frac{n}{m} = n$$

again a contradiction. The lower bound is also sharp whenever m is a factor of n by allowing $\lambda_i = \frac{n}{m}$ for all $1 \leq i \leq m$. We can now use these bounds to prove the inequalities in (2).

We substitute $\lambda_1 \geq \frac{n}{m}$ in (3) to prove the first inequality:

$$\begin{aligned} \kappa(\lambda^*) &= \frac{\log_2(\lambda_1)}{\log_2(n)} \\ &\geq \frac{\log_2\left(\frac{n}{m}\right)}{\log_2(n)} = \frac{\log_2(n) - \log_2(m)}{\log_2(n)} \\ &= 1 - \kappa(\lambda). \end{aligned}$$

Next, we can substitute $\lambda_1 \leq n - m + 1$ again into (3) to prove the second inequality:

$$\begin{aligned} \kappa(\lambda^*) &= \frac{\log_2(\lambda_1)}{\log_2(n)} \\ &\leq \frac{\log_2(n - m + 1)}{\log_2(n)}. \end{aligned}$$

We cannot break up the log in this case, but we can write m as a function of $\kappa(\lambda)$ by inverting κ.

$$\begin{aligned} \kappa(\lambda) &= \frac{\log_2(m)}{\log_2(n)} \\ \log_2(n)\kappa(\lambda) &= \log_2(m) \\ 2^{\log_2(n)\kappa(\lambda)} &= 2^{\log_2(m)} \\ n^{\kappa(\lambda)} &= m \end{aligned}$$

Substituting this back in to finish our bound we indeed see that

$$\kappa(\lambda^*) \leq \frac{\log_2(n-m+1)}{\log_2(n)} = \frac{\log_2(n-n^{\kappa(\lambda)}+1)}{\log_2(n)}.$$

To see that these bounds are sharp we simply need to provide a $[\lambda, \lambda^*]$ pair for each inequality which makes it into an equality. For the upper bound this is quite easy. Given any $q < n$ we have a partition $\lambda = \langle q, 1, \ldots, 1\rangle$ and then $\lambda^* = \langle m, 1, \ldots, 1\rangle$. Since this achieves the upper bound on λ_1 that we used in proving the upper bound on $\kappa(\lambda^*)$ we can turn the one inequality into an equality. For the lower bound we do the same thing. First assume that n is not prime, so we can write it as $n = n_1 \cdot n_2$ for some $n_1, n_2 < n$ and $n_1, n_2 \in \mathbb{Z}$. Then we let $\lambda = \langle n_1, n_1, \ldots, n_1\rangle$ where we have n_2 copies of n_1, and $\lambda^* = \langle n_2, n_2, \ldots, n_2\rangle$ with n_1 copies of n_2. Again, because this achieves the lower bound on λ_1 in terms of m we can turn our inequality into an equality. Now, in the case that n is prime we still achieve our lower bound for the trivial partition $\lambda = n$ where $\kappa(\lambda) = 0$. Then $\lambda^* = \langle 1, 1, \ldots, 1\rangle$ and we have $\kappa(\lambda^*) = 1 = 1 - \kappa(\lambda)$.

In order to illustrate the bounds in Proposition 1 see Fig. 11. The x axis is $\kappa(\lambda)$ and the y axis is $\kappa(\lambda^*)$. The line and curve are the bounds, and you can see many instances of points sitting on the bounds. This picture is for $n = 20$.

Given that there are sharp bounds on $\kappa(\lambda^*)$ in terms of $\kappa(\lambda)$ we asked whether or not we can do the same for $G(\lambda^*)$. In fact we can get a lower bound, but the form is much more complicated. And an upper bound seems nonexistent from looking at the points themselves. In Fig. 12 we have a picture similar to that in Fig. 11, again

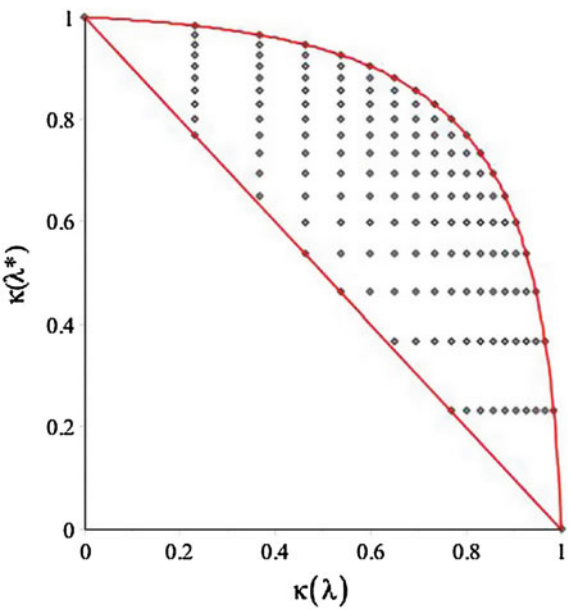

Fig. 11 Plot of $\kappa(\lambda^*)$ versus $x = \kappa(\lambda)$ for $n = 20$

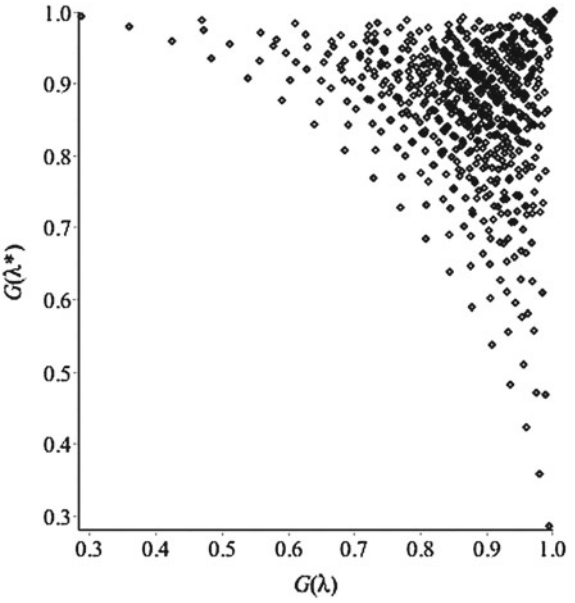

Fig. 12 Plot of $G(\lambda^*)$ versus $x = G(\lambda)$ for $n = 20$

for $n = 20$, but this time we give the G values. Notice that there is a clear lower bound curve traced out by the points. These lowest elements correspond to the case where $\lambda = \langle \lambda_1, 1, 1, \ldots, 1 \rangle$ and then $\lambda^* = \langle m, 1, 1, \ldots, 1 \rangle$ since partitions of this form have the lowest G value. However, the form of G is much more complicated than that of κ and we are not able to invert G to get a function for m in terms of $G(\lambda)$. Therefore, the solution is more of a numerical approximation than a closed form function.

6 Conclusion

In this paper we introduced two functions, dispersion (κ) and smoothness (G), which measure the shape of frequency distributions in two different ways. First, dispersion assesses how many bins, or labels, there are in the distribution versus how many objects. If there are a similar number of objects and bins then the distribution is very dispersed, but if there are many more objects than bins then the distribution is very narrow. Secondly, smoothness uses a normalized entropy to measure how close the distribution is to uniform. We showed how these measures function on directed labeled graphs and more specifically to the cyber security use case. Finally, we explored how G and κ change when we make specific changes to the partition through Young diagram manipulation.

We believe that these two functions can help discover changes in evolving data and characterize labeled directed multigraphs in much of the same way as the degree distribution characterizes unlabeled graphs.

Acknowledgments The research described in this paper is part of the Asymmetric Resilient Cybersecurity Initiative at Pacific Northwest National Laboratory. It was conducted under the Laboratory Directed Research and Development Program at PNNL, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy.

References

1. G. Birkhoff, *Lattice Theory*, vol. 25, 3rd edn. (American Mathematical Society, Providence, 1940)
2. G. de Cooman, D. Ruan, E. Kerre (eds.), *Foundations and Applications of Possibility Theory* (World Scientific, Singapore, 1995)
3. D. Dubois, H. Prade, *Possibility Theory* (Plenum Press, New York, 1988)
4. C. Joslyn, W. Cowley, E. Hogan, B. Olsen, Discrete mathematical approaches to graph-based traffic analysis, in *2014 International Workshop on Engineering Cyber Security and Resilience (ECSaR14)* (2014)
5. G. Klir, *Uncertainty and Information: Foundations of Generalized Information Theory* (Wiley, Hoboken, 2006)
6. R.P. Stanley, *Enumerative Combinatorics*, vol. 1 (Cambridge UP, Cambridge, 1997)
7. Visual Analytics Science and Technology (VAST) Challenge (2013). http://vacommunity.org/VAST+Challenge+2013
8. O. Wolkenhauer, *Possibility Theory with Applications to Data Analysis* (Wiley, New York, 1998)
9. G.M. Ziegler, On the poset of partitions of an integer. J. Comb. Theory, Ser. A **42**(2), 215–222 (1986)