Preprint typeset using LaTeX style emulateapj v. 12/16/11

AN INFORMATION THEORETIC FRAMEWORK FOR CLASSIFYING EXPOLANETARY SYSTEM ARCHITECTURES

Gregory J. Gilbert[1] & Daniel C. Fabrycky[1]

ABSTRACT

We use information theory to describe the architectures of exoplanetary systems.

Keywords: planetary systems; other keywords

1. INTRODUCTION

Describing exoplanetary system architectures is an inherently complex task. Although the basic observables for each individual planet (mass, radius, period, inclination, and eccentricity) are relatively straightforward to measure and compare, there is no single preferred way to combine these variables in order to assess system-level trends. Compounding the problem, the number of possible combinations of parameters scales rapidly with the number of planets in a system. Furthermore it is not always clear how to compare, say, a 2-planet system to a 6-planet system.

The standard approach has been to consider pairwise statistics, typically of adjacent planet pairs. For example, one might compute the ratios of planet masses, radii, orbital periods, or velocity-scaled transit durations. This approach has been quite fruitful, leading to powerful constraints on the distributions of planet sizes and orbital arrangements in multiplanet systems (Lissauer et al. 2011; Ciardi et al. 2013; Fabrycky et al. 2014; Winn & Fabrycky 2015; Millholland et al. 2017; Wang 2017; Weiss et al. 2018). However, this pairwise approach has two major drawbacks. First, this approach implicitly assumes that a system's global architecture is well described by pairwise statistics, which may or may not be the case. In reality, complex systems often display emergent properties that arise from interactions between independent components. Second, employing pairwise ratios does little to reduce the dimensionality of the problem at hand, and if non-adjacent planet pairs are considered actually *increases* the dimensionality of the problem.

A better approach would be to consider all planets in a system simultaneously using higher-order statistics that reduce the number of dimensions to a tractable level. Consider, for example, planet masses. Rather than tallying N masses and $\mathcal{O}(N^2/2)$ mass ratios, we aim to summarize the partitioning of mass between planets as a single number. For a 5-planet system, our approach would reduce the dimensionality by four, whereas the standard approach would increase the dimensionality by ten if all masses and mass ratios were considered independently. The benefits will be at least as substantial for any other quantity considered.

Our goal then is this: we aim to identify a small number of parameters which parsimoniously capture the global architecture of an exoplanetary system.

Beyond mathematical convenience, there is good physical motivation for prioritizing the system over the individual planets. Because all planets in a system share a common formation history, their properties are intrinsically linked, and thus planets do not constitute independent samples. A corollary is that any population level studies that treat planets independently will be inherently biased by correlations between the input variables. Stars, however - and by extension systems - do constitute independent samples, at least to first order when ignoring the effects of cluster environment or stellar multiplicity on planet formation. We therefore argue for a subtle yet radical shift in perspective: rather than treating the *planet* as the fundamental unit of exoplanet science, we treat the *system* as the fundamental unit.

But how do we reduce a complex system of planets to just a few numbers? Fortunately, there is a branch of mathematics well-suited to the task, aptly named "complexity theory," which is itself an extension of information theory. In this context, the word "complex" refers to any system which has distinct properties that arise from the interactions between components. Under this definition, planetary systems are indeed complex, and so we believe that complexity theory is the right tool for the job.

This paper is organized as follows. In §2 we review some relevant ideas from the information theory literature. In §3 we propose several new measures for classifying exoplanetary systems. In §4 we combine these measures, search for clusters of distinct system type. In §5 we compare real *Kepler* systems to synthetic populations. In section §6 we investigate whether the observed trends can be explained a subpopulation of systems with as-yet undetected planets. In §7 we summarize our results discuss possible modifications and future additions to our classification scheme.

2. OVERVIEW OF INFORMATION THEORY

Because the application of information theory is relatively novel to the field of astrophysics in general, and to the study of exoplanets in particular, here we review some foundational ideas from the information theory literature. The majority of this review is borrowed from Lopez-Ruiz, Mancini, & Calbet (2010), and so we refer the interested reader there for further detail.

2.1. *The big idea: convex complexity*

The seminal work of information theory came from Claude Shannon (1948), who linked the information content of a system to its entropy via the now-famous "Shannon information," or "Shannon entropy," defined as

Email: gjgilbert@uchicago.edu
[1] Department of Astronomy and Astrophysics, University of Chicago, 5640 S. Ellis Ave., Chicago, IL 60637, USA

$$H \equiv -\sum_{i=1}^{N} p_i \log p_i \quad (1)$$

where H is the Shannon information and p_i are occupancy probabilities for the N possible states of the system[2] being studied. The units of entropy are variously taken to be bits (for $\log_2$), nats ($\log_e$) or bans ($\log_{10}$). Although system behavior is identical regardless of units, there is no single standard for which base to use, and so one should take care when comparing results, especially when doing so across scientific disciplines. Unless otherwise noted, we use natural logarithms throughout this work.

In order to gain an intuition for how entropy behaves, it is instructive to examine two end-member cases: an ideal gas (maximum entropy) and a perfect crystal (zero entropy). For an ideal gas, each energy state has equal probability, and so Equation 1 is maximized, whereas for a perfect crystal a single energy state has $p_i \to 1$, and so Equation 1 is minimized. From the perspective of information theory, one can interpret this result as follows. The information content required to describe a perfect crystal is minimal - one needs only a single number (or perhaps a small handful of numbers) to specify the lattice bond length(s) in order to capture the entire structure of the crystal. For an ideal gas, on the other hand, one must specify the position and momentum of every individual particle in order to describe the full structure. The solution from statistical mechanics, describing the gas as a distribution over momentum states, is exactly what its name implies - statistical - and does not truly capture the full information content of the physical system.

The mind may rebel at the notion that an ideal gas holds maximal information content. "Obviously" the entropy of the system is at a maximum, yet at the same time it is "obvious" that little useful information can be retrieved from the individual particles. For a fully randomized system, even though the formal information content (i.e. the entropy) has been maximized, the extractable information has been minimized. Clearly, entropy alone is not sufficient to capture the behavior of these systems.

A complementary statistic to entropy is the *disequilibrium*, defined as

$$D \equiv \sum_{i=1}^{N} (p_i - \frac{1}{N})^2 \quad (2)$$

For our end-member cases, D is minimized for the ideal gas, in which every energy state has equal probability, and maximized for the perfect crystal, in which a single energy states dominates completely. Thus, disequilibrium shows qualitatively inverse behavior to entropy.

Our intuition tells us that both an ideal gas (fully randomized) and a perfect crystal (fully ordered) have zero complexity. But zero complexity is **not** synonymous with zero entropy; a zero entropy system will have zero complexity, but zero complexity system need not have zero entropy. Because entropy is often colloquially described as chaos or disorder, the distinction between complexity and entropy can sometimes become blurred. So, at the risk of redundancy, we reiterate that our aim is to distinguish between systems with low versus high complexity, which may not always be the same as distinguishing between systems with low versus high entropy.

[2] Here the word "system" is used in the general sense of any physical system, and not in the particular sense of a planetary system.

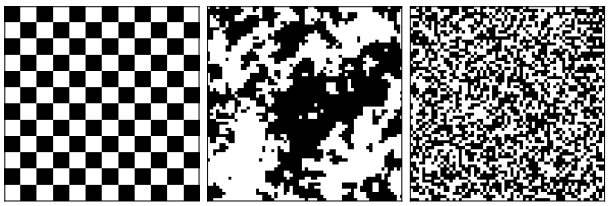

Figure 1. An illustration of the intuitive notion of complexity. Both complete order (left panel) and total randomness (right panel) have low complexity, while the intermediate state (middle panel) displays patterns that we might intuitively call complex.

What we desire then is a measure of complexity which goes to zero for both perfectly ordered and perfectly random systems, peaking at some maximum complexity for an intermediate state where a balance between entropy and equilibrium is achieved. An example of such a situation is shown in Figure 1. Using information theory parlance, measures which meet the above criteria are sometimes called *convex* complexity measures. Our discussion of information theory has so far implicitly treated convex complexity measures as if these are the *only* way to define complexity. In fact, some definitions of complexity are defined as monotonic functions of entropy, in direct contradiction to our arguments above. However, these definitions tend to arise primarily in the field of computer science (cf. Chaitin 1966; Kolmogorov 1968). In the social, biological, and physical sciences, convex complexity measures dominate, and so as not to confuse matters, we will restrict ourselves to consideration of convex measures.

There are two convex complexity measures commonly used in the physical sciences. The first was introduced by López-Ruiz, Mancini, & Calbet (1995), who defined complexity as the product of entropy and disequilibrium:

$$C \equiv H \cdot D = -K \left(\sum_{i=1}^{N} p_i \log p_i \right) \cdot \left(\sum_{i=1}^{N} (p_i - \frac{1}{N})^2 \right) \quad (3)$$

where K is a positive, real constant. C is a statistical measure of complexity and is sometimes called LMC complexity (after the authors; no relation to the Large Magellanic Cloud). A schematic representation of the behavior of H, D, and C is shown in Figure 2.

The second measure, a quantity closely related LMC's *statistical* measure of complexity, is the *simple* measure of complexity proposed by Shiner, Davison, & Landsberg (1999). Both measures have the same qualitative feature of going to zero either near complete order or near total randomness, peaking at a maximum complexity somewhere in the middle. The so-called SDL complexity, Γ, is defined as

$$\Gamma \equiv \eta^{\alpha}(1-\eta)^{\beta} \quad (4)$$

$$\eta = H/H_{max} \quad (5)$$

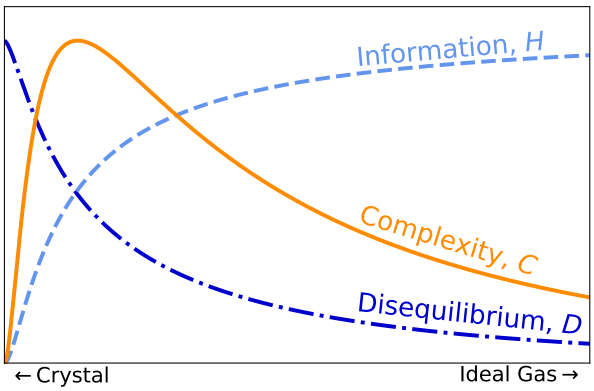

Figure 2. A schematic representation of the behavior of information, H, disequilibrium, D, and LMC complexity, $C = H \cdot D$. Complexity reaches a maximum at a midpoint between complete order (perfect crystal) and total randomness (ideal gas). Any measure of convex complexity must satisfy these boundary conditions. Adapted from Lopez-Ruiz, Mancini, & Calbet (2010).

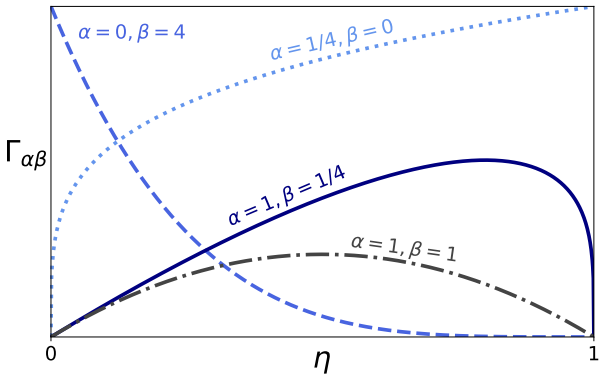

Figure 3. SDL complexity, Γ, vs disorder, η. Setting either α or β equal to zero results in a monotonic function $\Gamma(\eta)$, whereas setting both α and β to positive, nonvanishing values produces a convex complexity curve with varying degrees of skewness. See Equations 4 & 5 for definitions of quantities. Adapted from Shiner, Davison, & Landsberg (1999).

where α and β parameterize the relative weighting of disorder, η, vs order, $1 - \eta$, and H_{max} is the maximum entropy achievable by the system being studied. Figure 3 shows how Γ behaves for a few values of α and β. Note that certain choices of α or β can also produce a complexity curve in which Γ is a monotonic function of entropy (i.e. not convex). For an in-depth comparison of C and Γ, see Panos et al. (2007), who explore the behavior of these two measures when applied to atomic structure.

When applying the idea of convex complexity to exoplanetary system architectures, we opt to use C over Γ for two reasons. First, Γ derives directly from the entropy, without explicit consideration of disequilibrium. Because a dominant feature of exoplanet architectures appears to be their frequent nearness to equipartitioning, we believe that C is a more appropriate choice for characterizing complexity. Second, whereas calculating C - to within a normalization factor - provides a fixed value dependent only on the number and occupancy probabilities of allowed states, calculating Γ includes two additional parameters, α and β, which allows for a higher level of modeling flexibility that we do not believe is warranted by the current quality of available data. However, we remain open to the idea that Γ and the framework of SDL complexity (or even some other complexity scheme not discussed here) may ultimately prove to be the superior measure for characterizing exoplanetary system architectures.

As an aside, it is worth noting that all of our above discussions have been limited to a broad class of ideas that fall under the umbrella of *algorithmic* complexity. These measures all derive in one form or another from Shannon entropy. Roughly speaking, there are two other main branches of complexity theory: *deterministic* complexity, commonly known as chaos theory, and *aggregate* complexity, which focuses on how individual elements create complex patterns and systems, such as those found in the transmission of epidemic diseases or in the migration patterns of birds. The distinctions between these three categories are somewhat arbitrary and sometimes ill-defined, and each branch shares many overlapping ideas with the other two. Nevertheless, such distinctions can be useful to make when diving into the vast literature of information theory. We direct the reader to three excellent reviews by May (1976), Manson (2001), and Lansing (2003) for further exploration of these ideas.

2.2. *Application to astrophysics*

Before defining our new measures to describe exoplanetary system architectures, we pause briefly to outline our philosophical approach to this problem.

Our technique is descriptive and not tied to any underlying theories of planet formation or orbital dynamics. Why? Because we wish to characterize patterns in the data without introducing biases that might arise from making physical assumptions. Our philosophy is "describe first, explain later."

As a motivating example, consider Edwin Hubble's development of a morphological classification scheme for galaxies (Figure 4; Hubble 1926, 1936). To quote Hubble directly, while developing this scheme, "deliberate effort was made to find a descriptive classification which should be entirely independent of theoretical considerations...the basis of the classification is descriptive and entirely independent of any theory." Even though Hubble's physical explanation was precisely *backwards*, the sequence nonetheless captured many of the most important features of galaxy structure. Indeed, Hubble's purely morphological description was so successful that it remains relevant nearly a century later!

We adopt a similar approach here, eschewing the theoretical framework of planet formation in favor of a purely descriptive technique. If our methods are sound, they should naturally capture the important features of the population of systems. In order to be accessible to the community, we strive to make our measures straightforward to calculate and easy to interpret. [CONCLUDING SENTENCE].

3. DESCRIPTION OF OUR CLASSIFICATION SCHEME

Here we propose several measures to quantify the global structure of planetary systems. Some are modifications of quantities which already exist in the literature, while others are new as of this work. Our proposed measures are as follows:

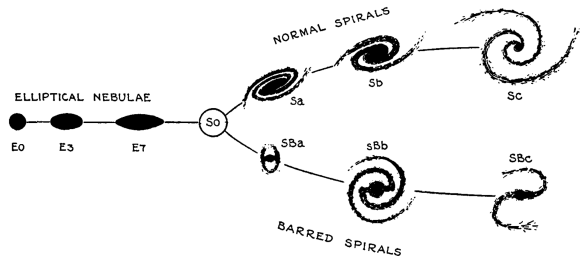

Figure 4. Hubble's "tuning fork" classification of galaxies. Even though Hubble got the explanation precisely *backwards*, his morphological classification scheme nonetheless captured some of the most important features of galaxy evolution and effectively distinguished galaxies into physically meaningful groups. This classification was purely morphological and even grouped galaxies by complexity. We seek to take a similar approach to classifying exoplanet systems. Figure reproduced from Hubble (1936)

1. *Dynamical mass*, μ, sets the overall mass scale of the system;
2. *Mass partitioning*, $\mathcal{Q}$, captures the variance in masses between planets;
3. *Monotonicity*, $\mathcal{M}$, describes the size ordering of the planets;
4. *Characteristic spacing*, $\mathcal{S}$, is the average separation between planets in mutual Hill radii;
5. *Gap complexity*, $\mathcal{C}$, summarizes the relationships between orbital periods;
6. *Flatness*, f, is related to the scatter in mutual inclinations;
7. *Multiplicity*, N, is the observed number of planets in a system.

When evaluating these measures on real systems, we use the catalogue from the California Kepler Survey (CKS; Johnson et al. 2017; Petigura et al. 2017). To ensure a high-quality sample, we cross-match all candidate planets with *Kepler* Data Release 25 (DR25; Thompson et al. 2018). We next make a few reasonable cuts to remove false positives, grazing transits ($b > 1 - r_p/R_\star$) and low signal-to-noise ($SNR < 7.1$) objects. Finally, in order to ensure that all stellar characterization is accurate, we apply restrictions on stellar radius, temperature, "isochrone parallax," and dilution following the procedures described in section 4.2 of Fulton & Petigura (2018) and using the stellar companion catalogue of Furlan et al. (2017). After applying these cuts, we are left with 864 planets in 335 multiplanet systems. Of these, 452 planets are found in 129 high-multiplicity ($N \geq 3$) systems. All of the stars in the sample have been well characterized by Gaia (Gaia Collaboration et al. 2016) and CKS spectroscopy, which provide tight constraints on stellar density, thus allowing us to more precisely determine transit durations and, by extension, mutual inclinations. Furthermore, the high precision on stellar parameters translates into correspondingly high precision on planetary periods and radii.

Because masses are not available for the majority of *Kepler* planets, we convert radii to mass using the probabilistic mass-radius relations of [NEIL & ROGERS 2019]. These relations are similar to the probabilistic forecasting of Chen & Kipping (2017) but also incorporate information on orbital period and stellar insolation. Although there is considerable scatter in the mass-radius relation, with a sample of several hundred planets we expect this scatter to marginalize out so that we can still see statistical trends within the population of systems. We choose to work with mass rather than directly working with radius because mass is the more fundamental quantity tied to planet formation. In order to maintain a homogeneous sample, we do not use masses derived directly from radial velocities or transit timing variation measurements, even when such masses are available. These mass measurement are, however, incorporated into the probabilistic model of [NEIL & ROGERS].

Our decision to use a sample of only transiting planets is driven by a desire for convenient comparison with previous works (in particular Weiss et al. 2018). However, our techniques are not tied to any particular detection method, and so the measures described in this section could be easily generalized to a population of planets detected by any other method or even some heterogeneous combination of detection methods. To keep things simple, we stick to a homogeneous CKS catalogue and save the application of these measures to other exoplanet populations for future work.

3.1. *Dynamical mass, μ*

Mass is arguably the most fundamental property of an individual planet, and so describing the mass scale of each system a natural place to start when characterizing exoplanetary system architectures. The **dynamical mass** of the system is defined as

$$\mu \equiv \sum_{i=1}^{N} m_i / M_\star \tag{6}$$

where m_i are planet masses and $M_\star$ is the stellar mass, with the term "dynamical mass" taken following Jontof-Hutter et al. (2016). We choose to report mass as the system-to-star mass ratio rather than the simpler total integrated mass because planet formation and orbital dynamics are more closely related to disk-to-star and planet-to-star mass ratios than to total mass. Indeed, we note that the distribution of dynamical masses is conspicuously peaked near $\mu \approx 10^{-4}$ (Figure 6), commensurate with the common dynamical masses of the Jovian, Saturnian, and Uranian moon systems (e.g. Mosqueira & Estrada 2003; Canup & Ward 2006), hinting at a common formation pathway for exoplanet systems and giant planet satellites (Chiang & Laughlin 2013; Miguel et al. 2019).

The cumulative density function (cdf) of μ and μ/N for 2, 3, and 4+ planet systems are shown in Figure 5. We find that both the Kolomogorov-Smirnov test statistic (K-S test) and Anderson-Darling test statistic (A-D test) indicate that while $\log \mu$ is drawn from different distributions for different multiplicities, when μ is normalized by multiplicity all systems appear to be drawn from the same underlying distribution. [IS THIS IN CONTRAST TO STEFFEN OR WEISS?]. The straightforward interpretation of this trend is that the average planet size (relative to host star) is the same for all multiplicities

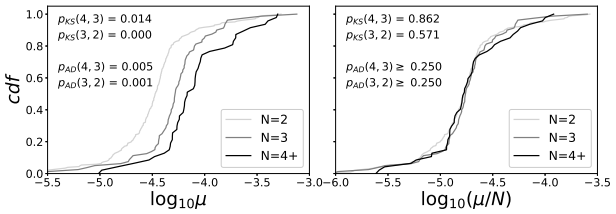

Figure 5. Cumulative density function of $\log_{10}\mu$ (left panel) and $\log_{10}(\mu/N)$ (right panel). Both Kolmogorov-Smirnov and Anderson-Darling tests indicate that while $\log\mu$ is drawn from different distributions for different multiplicities, when μ is normalized by multiplicity all systems appear to have an average planet size drawn from the same underlying distribution.

$N \geq 2$. We further hypothesize that these dynamical mass variations between multiplicities indicate that many of the lower multiplicity systems host additional undetected planets.

3.2. *Mass partitioning, $\mathcal{Q}$*

We define the **mass partitioning** of a system as

$$\mathcal{Q} \equiv \left(\frac{N}{N-1}\right) \cdot \left(\sum_{i=1}^{N}(m_i^* - \frac{1}{N})^2\right) \tag{7}$$

$$m_i^* = m_i / \sum_i^N m_i \tag{8}$$

The second bracketed term in Equation 7 is simply the disequilibrium, D, calculated from Equation 2 with the substitution $p_i \rightarrow m_i^*$, where m_i^* is the normalized planet mass defined following Equation 8. In the language of occupancy probabilities, m_i^* can be thought of as the probability that an infinitesimal mass element dm resides in a particular planet. In simple terms, m_i^* is the fraction of total system mass (excluding the star) contained in an individual planet. The prefactor $N/(N-1)$ normalizes $\mathcal{Q}$ to the range (0,1). Thus, any system with all equal mass planets will have $\mathcal{Q} = 0$ while a system with one dominant giant planet and $N-1$ tiny planets will have $\mathcal{Q} \rightarrow 1$.

We note that our definition of mass partitioning is closely to the intra-system mass dispersion measures of Millholland et al. (2017) and Wang (2017). Our present work was largely inspired by these two studies, an so we are indebted to their efforts. The critical advantage of our method over these antecedent works is that our measure is more intuitive to interpret and is more explicitly linked to the global architecture of each system, which facilitates not only intra-system comparison by also inter-system comparison.

The most striking feature of the distribution of $\mathcal{Q}$ (Figure 6) is that nearly all systems show a high degree of uniformity of planet sizes, i.e. low $\mathcal{Q}$, confirming previous results (Millholland et al. 2017; Wang 2017; Weiss et al. 2018). There is no strong correlation between dynamical mass, μ and mass partitioning, $\mathcal{Q}$, although for relatively massive systems ($\log_{10}\mu \gtrsim -3.8$) there is a positive correlation between $\mathcal{Q}$ and $\log_{10}\mu$, a result which is expected under the runaway gas accretion model of giant planet formation.

We find no obvious difference in the $\mathcal{Q}$ distribution of 3 planet vs 4+ planet systems as based on the results of K-S and A-D tests, suggesting that regardless of multiplicity the planets in these systems are drawn from the same underlying $\mathcal{Q}$ distribution. We find marginal evidence ($p_{KS} = 0.084, p_{AD} = 0.045$) that 2-planet systems may be drawn from a different population than higher multiplicity systems. These marginal test statistics would be expected if some, but not all of the 2-planet systems belong to the same underlying physical population as the high multiplicity systems. However, because some of our other complexity measures are undefined for systems with fewer than 3 planets, we do not investigate this point further.

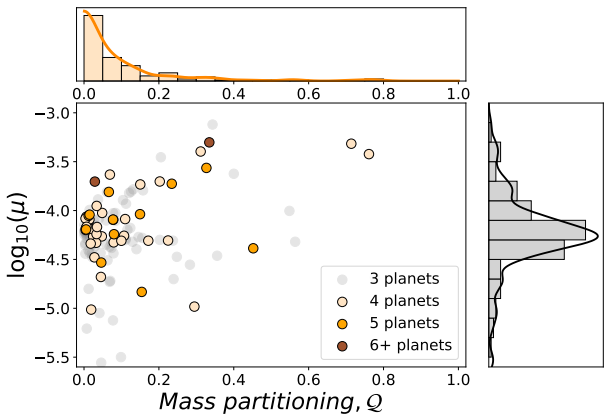

Figure 6. Distribution of mass partitioning, $\mathcal{Q}$, vs. dynamical mass μ. We find that the majority of systems have low $\mathcal{Q}$, indicating that planets within a system tend to be the same size. Among the few systems with large $\mathcal{Q}$, many also have high values of μ, which is to be expected based on the definition of $\mathcal{Q}$. Conversely, we find few planets with large $\mathcal{Q}$ but low μ, indicating that if giant planets do form, runaway gas accretion occurs unequally between planets in a system.

3.3. *Monotonicity, $\mathcal{M}$*

Several prior studies have found that exoplanets are preferentially arranged with larger planets exterior to smaller planets (Ciardi et al. 2013; Millholland et al. 2017; Weiss et al. 2018; Kipping 2018), although there has recently been some controversy surrounding this claim (Zhu 2019).

In order to capture the degree to which planets are ordered by mass, we define the **monotonicity** as

$$\mathcal{M} \equiv \rho_S \mathcal{Q}^{1/N} \tag{9}$$

where ρ_S is the Spearman rank-order coefficient calculated using planet masses as the input variables. The rank-order takes a value $\rho_S = 1$ for perfectly positive monotonic systems, $\rho_S = -1$ for perfectly negative monotonic systems, and $\rho_S = 0$ for systems with no evidence of monotonic behavior. We include the multiplicative factor $\mathcal{Q}^{1/N}$ because while rank-order captures the degree to which a sequence is monotonic, it provides no information regarding the magnitude of any such monotonic trend. Multiplying by $\mathcal{Q}$ downweights $\mathcal{M}$ towards zero for any systems in which the planets are close to evenly sized and thus for which any evidence of monotonicity is more likely to be due to statistical noise. The factor $1/N$ scales this $\mathcal{Q}$ factor so that a high multiplicity system will have its evidence of monotonicity preserved even if there is little variation in planet masses. Using ρ_S directly would also be inappropriate because it can only

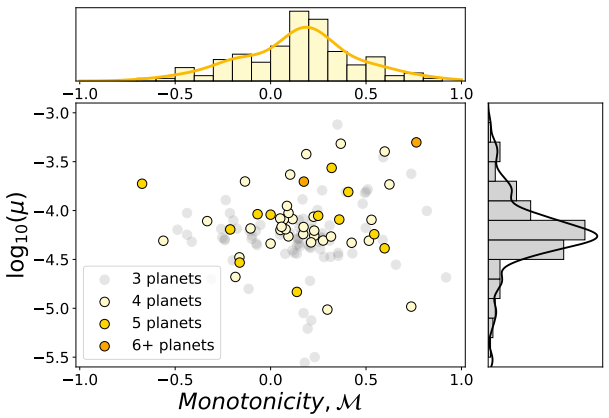

Figure 7. Distribution of monotonicity, $\mathcal{M}$, vs. dynamical mass μ. The distribution of $\mathcal{M}$ is peaked at small positive values, inditing that there may be preference for planets to be arrange with larger bodies exterior to smaller bodies. At present, it is unclear whether this trend in monotonicity is physical in nature of the result of observational biases.

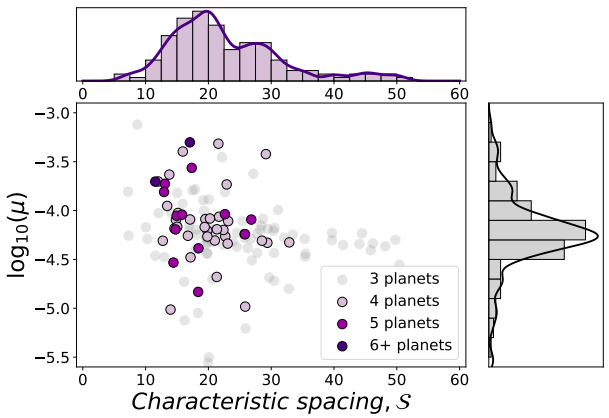

Figure 8. Distribution of characteristic spacing, $\mathcal{S}$, vs. dynamical mass μ. The distribution is peaked at § $\approx$ 20 mutual Hill radii, in agreement with Weiss et al. (2018). We also find a tentative secondary peak near $S \approx 30$. This "echo" peak would be expected if there is a large population of intrinsic 4-planet systems in which one of the intermediate planets has not yet been detected.

take a few discrete values for small N and thus is not on its own flexible enough to describe the apparent size-ordering of planets. Note that $\mathcal{M}$ is normalized to the range (-1,1) and has the same qualitative interpretation as ρ_S.

We find that 72% of high multiplicity ($N \geq 3$) systems have $\mathcal{M} > 0$, strengthening the finding of Weiss et al. (2018) that 65% of planet pairs are ordered with the larger planet exterior to the smaller planet. In addition, while the smallest observed negative monotonicity value is $\mathcal{M} = -0.67$, there are five systems which have at least equally strong positive monotonicity $\mathcal{M} > +0.67$. Similarly, although there are only two systems with $\mathcal{M} < -0.5$, there are sixteen systems with $\mathcal{M} > +0.5$. In section 5 of this paper we investigate whether this apparent size ordering in physical in nature or the result of observational biases. The distribution of $\mathcal{M}$ is shown in Figure 7.

3.4. *Characteristic spacing, $\mathcal{S}$*

To describe the orbital spacing of a system, we define the **characteristic spacing**, $\mathcal{S}$ as the average separation between planets in units of mutual Hill radii

$$\mathcal{S} \equiv mean(\Delta_H) \tag{10}$$

where the scaled separation Δ_H and mutual Hill radius, r_H, for each adjacent planet pair are calculated from

$$\Delta_H = (a' - a)/r_H \tag{11}$$

$$r_H = \left(\frac{m' + m}{3M_\star}\right)^{1/3} \left(\frac{a' + a}{2}\right) \tag{12}$$

and a is the semimajor axis of each planet. Primed variables are outer planets in an adjacent pair and unprimed variables are inner planets.

Figure 8 shows the behavior of $\mathcal{S}$ vs $\log \mu$. We reproduce previous findings (Fang & Margot 2012; Pu & Wu 2015; Dawson et al. 2016; Weiss et al. 2018) that planets are separated by $\sim$20 mutual Hill radii. We also identify a tentative secondary peak at $\mathcal{S} \approx 28$. This "echo peak" would be expected near $\mathcal{S} \approx 30$ if there exists a significant population of evenly-spaced 3-planet systems which are intrinsically four planet systems in which one of the two intermediate planets has not been detected. We discuss this hypothesized subpopulation further in our discussion of gap complexity and in Section 6 below.

3.5. *Gap complexity, $\mathcal{C}$*

We now define a measure which we term the **gap complexity**, following Equation 3 and restated here as

$$\mathcal{C} = -K \left(\sum_{i=1}^{n} p_i^* \log p_i^*\right) \cdot \left(\sum_{i=1}^{n} (p_i^* - \frac{1}{n})^2\right) \tag{13}$$

$$p_i^* \equiv \frac{\log(P'/P)}{\log(P_{max}/P_{min})} \tag{14}$$

where K is a normalization constant, $n = N - 1$ is the number of gaps between planets and p_i^* are pseudo-probabilities computed from the planets' orbital periods using Equation 14. Here, P' is the orbital period of the outer planet in an adjacent pair, P is that of the the inner planet, P_{max} is the maximum period of any planet in the system, and P_{min} is the minimum period. In this way p_i^* are automatically normalized so that $\sum p_i^* = 1$.

Because orbital periods are not so easily converted to probabilities as masses are, gap complexity should be interpreted as a purely descriptive term and not as a traditional physical quantity. We do note, however, that orbital periods could in principle be transformed into specific orbital energies, which could then be re-expresed as occupancy state probabilities. Nevertheless, these extra steps would not necessarily result in a better outcome, so in order to keep things simple and maintain focus on the "big-picture" of our methodology, we opt to work directly with orbital periods here.

The normalization constant, K, is chosen so that $\mathcal{C}$ is always in the range (0,1). A system with planets evenly spaced in log-period will have $\mathcal{C} = 0$, while $\mathcal{C} \to 1$ when the maximum $p_i^* \approx 2/3$ (Anteneodo & Plastino 1996); the exact value depends on the number of planets in the system. K can equivalently be expressed as $1/\mathcal{C}_{max}$,

Table 1
Numerically determined values of $\mathcal{C}_{max}$ vs n. Note that the values shown are for the number of gaps, n, rather than the number of planets, $N = n + 1$, in a system

n	$\mathcal{C}_{max}$
2	0.106
3	0.212
4	0.291
5	0.350
6	0.398
7	0.437
8	0.469
9	0.497

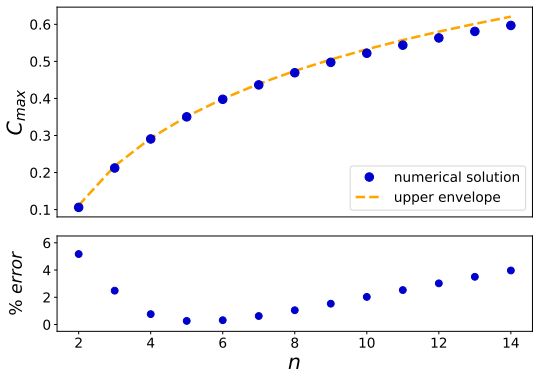

Figure 9. Numerically computed values of $\mathcal{C}_{max}$ for $n \leq 14$. The numerical values are well approximated by an upper envelope $\mathcal{C}_{max} \approx 0.262 \ln(0.766n)$, which is accurate to within 6% for the values of n shown. Using an upper envelope ensures that when this approximation is used $\mathcal{C}_{max}$ will remain in the range (0,1).

where $\mathcal{C}_{max}$ is the maximum possible complexity for a given n. Anteneodo & Plastino (1996) derive equations to numerically determine $\mathcal{C}_{max}$, and we present values for $n \leq 9$, corresponding to multiplicities for all known planetary systems, in Table 1. Alternatively, $\mathcal{C}_{max}$ can be approximated from the relation

$$\mathcal{C}_{max} \approx 0.262 \ln(0.766n) \tag{15}$$

The above relation is determined by fitting a power law to the numerically determined maximum complexity, $\mathcal{C}_{max}$ vs n for $n \leq 9$. The power law fit is then shifted to establish an upper envelope so for $\mathcal{C}_{max}$ so that after normalization $\mathcal{C}$ remains in the range (0,1). We stress that Equation 15 is provided merely for convenience and is an empirical quantity. In practice, using equation 15 amounts to less than 2.5% error for systems with $4 \leq N \leq 10$ and less than 6.1% error for all systems with $N \leq 15$ (Figure 9). A python script for calculating $\mathcal{C}_{max}$ by numerically solving the equations of Anteneodo & Plastino (1996) will be made available on github (https://github.com/gjgilbert/maiasaurus).

Having defined our gap complexity measure, one might reasonably ask why not simply use D or H directly to characterize the spacing between planets? H can be easily ruled out because planets in most real systems are roughly evenly spaced, and so are near maximum entropy, whether using p^* or even if using P'/P directly. So, H is of little practical use when comparing real system spacings. The choice of $\mathcal{C}$ over D is admittedly something of a subjective choice. We opt for $\mathcal{C}$ for two main reasons. First, for the population of observed systems, there is little actual dynamical range in D. In other words, D is relatively insensitive to the scale of variations in period spacings for real planetary systems. Second, we find that using $\mathcal{C}$ more closely matches our intuitive sense of which systems are complex and which are simple; the straightforward quadratic dispersion relation of D does not adequately capture the variety of system architectures observed by *Kepler* .

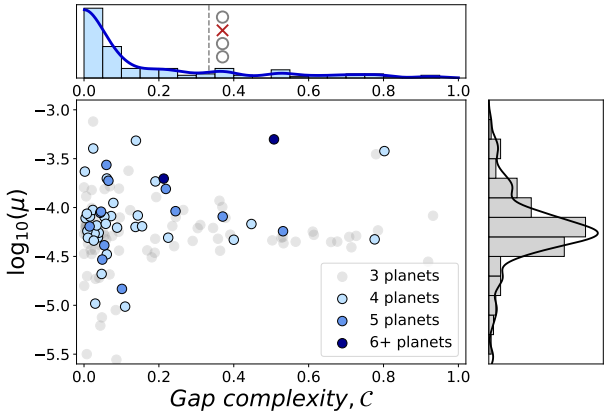

Figure 10. Distribution of gap complexity $\mathcal{C}$, vs. dynamical mass μ. The majority of systems have $C \lesssim 0.2$, indicating that planets tend to be evenly spaced in log-period within a given system. There is a heavy tail extending out towards high values of $\mathcal{C}$. The majority of systems in this tail have dynamical masses consistent with the population mean, suggesting that this tail may be dominated by systems in which an intermediate planet has not yet been detected.

Astrophysicists, accustomed as we are to quantities derived from first principles, may feel some understandable discomfort with the subjective nature of this approach. However, a subjective approach does not imply an unprincipled approach. Any classification scheme will contain a degree of subjectivity, even if the only subjective choice being made is which features of the system to consider. [DOES THIS POINT NEED TO BE DEFENDED FURTHER?]

Like $\mathcal{Q}$, the distribution of $\mathcal{C}$ is peaked near zero, indicating that most systems are uniformly spaced, confirming the other primary result of Weiss et al. (2018). In fact, when compared to population synthesis models, the system-level variable $\mathcal{C}$ indicates that planets within a system are even more regularly spaced than was previously inferred from pair statistics alone (see Section 5). In addition to the peak at $\mathcal{C} = 0$, we find a long, heavy tail containing some $\sim 25\%$ of systems extending out to high gap complexities. One possible explanation for this high-complexity tail is that there exists a significant subpopulation of systems which host additional undetected planets at periods intermediate to the known planets. We explore this hypothesis in detail in Section 6.

3.6. *Flatness, f*

To complete our classification scheme, we define a measure of **flatness**, f that describes how close a system comes to that predicted for a completely "cold" architecture with circular and coplanar orbits. If such a system were completely edge-on to our line of sight, its transit durations would be perfectly predictable, equaling the orbital period times the ratio of the stellar diameter to the

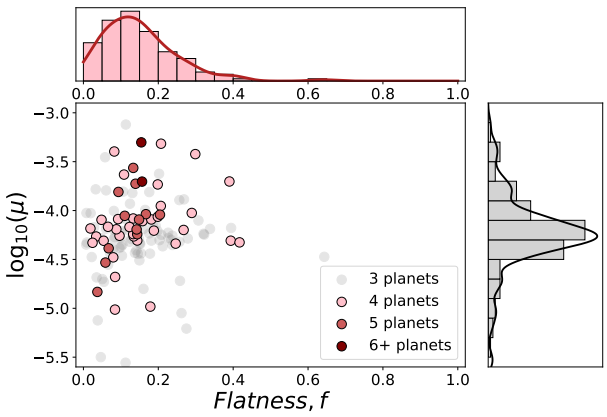

Figure 11. Distribution of flatness, f, vs. dynamical mass μ. The majority of systems have small but non-zero flatness, and we find no obvious correlation between f and μ.

orbital circumference. If such a system were not edge-on, each planet would cut a smaller chord across the star and hence have a shorter transit duration. Converting semi-major axis to period using Newton's version of Kepler's third law, we have for the T_1 to T_4 transit duration:

$$D = \left(\frac{3P}{G\rho_\star \pi^2}\right)^{1/3} \left((1+r)^2 - \left(\frac{G\rho_\star}{3\pi}\right)^{2/3} P^{4/3} \cos^2 i\right)^{1/2} \tag{16}$$

where $\rho_\star$ is the stellar mean density, $r = r_p/R_\star$, and i is the inclination from the sky plane. We fit this function using only the free parameter $\cos i$. The flatness measure f is the sum of the squared residuals remaining after the fit, divided by the sum of the squared D values before the fit.

We find that most systems are quite flat (f near zero), in agreement with Fabrycky et al. (2014). Although there is no hard upper boundary imposed on f, in practice only a pathological (and likely nonphysical) architecture could result in $f > 1$ and so f is therefore approximately normalized to the range (0,1). Based on K-S and A-D tests, we find no evidence that 3-planet systems are drawn from from a different underlying flatness distribution than higher multiplicity ($N \geq 4$) systems.

4. CLUSTERING AND CORRELATING

A primary benefit of our system-level approach is that we can now compare our measures against each other and search for correlations. Intuitively, one might expect that a system with equally-sized planets (low $\mathcal{Q}$) would also have planets which are equally-spaced (low $\mathcal{C}$) or tightly spaced (low $\mathcal{S}$, and a relatively coplanar geometry (low f). Interpreted through the lens of planet formation, it seems likely that a quiescent planet formation history that allows for equal mass partitioning might also allow for ordered period spacing and small mutual inclinations. Conversely, any chaotic stage during formation that disrupts orderly mass and energy partitioning would also be likely to excite inclinations.

In order to assess the strengths of relationships between our complexity measures, we employ the distance correlation metric (dCor; Szkely et al. 2007). Compared to the Pearson product-moment correlation coefficient, the distance correlation has the advantage of probing nonlinear relationships. Furthermore, dCor=0 only when two variables are independent.

We find that the strongest relationships exist between $\mathcal{Q}$ and $\log\mu$ (dCor=0.38) and between $\mathcal{C}$ and $\mathcal{S}$ (dCor=0.52). These strong correlations are to be expected, as they represent the covariance between our two mass measures ($\mathcal{Q}, \mu$) and two spacing measures ($\mathcal{C}, \mathcal{S}$), respectively. We also find strong correlation (dCor=0.48) between $\mathcal{Q}$ and $\mathcal{M}$, which is expected from the definition of $\mathcal{M}$. The correlation between most other variable pairs is weak but nonzero ($0.1 <$ dCor < 0.2), except for between $\mathcal{S}$ and f (dCor=0.36, $p < 2 \times 10^{-4}$). These two variables are positively correlated, indicating that the more tightly spaced systems are also flatter. Surprisingly, we do not find a commensurate correlation between $\mathcal{C}$ and f, but we do note that there is an observational towards low f. If a system has high mutual inclinations such that a planet is tilted off the limb of the star and does not transit, our inferred value of f would be lowered. Hence, if the more unevenly spaced systems are indeed less flat, this effect would be difficult to observe.

Even though the two-variable correlations between most parameter pairs are weak, it may still be possible to uncover hidden structure by considering all variables simultaneously. [MORE ON DIMENSIONALITY REDUCTION AND IDENTIFYING SUBPOPULATIONS]

A similar endeavor has been pursued at least once before by Alibert (2019) who used t-distributed stochastic embedding (T-SNE) to automatically reduced the dimensionality of systems in (R, P) space. Our approach is complementary but has the advantage of searching for clusters in a more intuitive parameter space using quantities that are explicitly tied to each system's global architecture. The main difference is that T-SNE is an unsupervised machine learning approach, whereas our method is very distinctly hands-on. The choice of which to use may come down to a matter of personal preference. We also note that T-SNE must also define a distance metric, so even though it may appear more objective at first there is the same degree of subjectivity at the outset.

We explore clustering in n-dimensional space using robust path-based spectral clustering (Chang & Yeung 2008) in order to quantify the similarity between systems and search for possible subpopulations of systems. In order to sufficiently sample an n-dimensional space, at least 2^n samples are needed. With 129 systems, we can therefore cluster in up to 7 dimensions. In practice, we use five: $\log\mu$, $\mathcal{Q}$, $\mathcal{S}$, $\mathcal{C}$, and f. We do not include $\mathcal{M}$ in our clustering searches because it is calculated explicitly from $\mathcal{Q}$ and is therefore correlated by definition. We do not include N because the discrete nature of this variable leads clustering algorithms to simply group systems by multiplicity.

Before proceeding, we apply multiplicative scaling factors to $\log_{10}\mu$ and $\mathcal{S}$ so that they are each approximately normalized to the range (0,1). This step weights each dimension equally in Euclidean distance space; $\mathcal{Q}$, $\mathcal{C}$, and f are already normalized to this range by definition. We explore a range of normalization factors and find that our results remain qualitatively unchanged for $\log\mu \rightarrow \frac{1}{k_\mu}\log_{10}\mu$ if $1.5 < k_\mu < 4$ and for $\mathcal{S} \rightarrow \mathcal{S}/k_S$ if $30 < k_S < 80$. For our final results, we use $k_\mu = 2$ and

$k_S = 40$.

The first step of standard spectral clustering is to construct a similarity matrix

$$s_{ij} = \begin{cases} e^{\|x_i - x_j\|^2/2\sigma^2} & i \neq j \\ 0 & i = j \end{cases} \quad (17)$$

which assigns a similarity value to each pair of points (i,j) based on their Euclidean distance in n-dimensional space. This matrix is mathematically equivalent to a graph representation in which each entry s_{ij} is the edge strength between nodes. Path-based clustering modifies the similarity matrix to account for strong but indirect paths between nodes in the graph (Chang & Yeung 2008). A critical advantage of path-based clustering over standard spectral clustering is that is that the path-based approach is insensitive to choice of scaling parameter σ which ordinarily can greatly impact the inferred clustering, yet is difficult to set in a self-consistent manner. The "robust" portion of robust path-based spectral clustering refers to an additional weight w_{ij} assigned to each pair of point such in order to account for local variation of cluster sizes and densities. In particular, the data are weighted to prioritize paths that pass either through dense regions or through sparse regions but penalize paths which leap between regions of varying density.

The common next step of all spectral clustering methods is to construct a graph Laplacian, $L = s_{ij} - d_{ij}$, where $d_{ij} = \sum_j s_{ij}$ along the diagonal and 0 otherwise. Clustering is then performed on the first k eigenvectors of L using the basic k-means algorithm. In order to select the number of clusters, k, to consider, we employ the eigengap heuristic, which states that the clustering solution is most stable when k is chosen such that the difference in eigenvalues is maximized (Ng et al. 2001; von Luxburg 2007). The first 8 eigenvalues for our clustering solution are shown in Figure 12. In tests with synthetic data, we find that compared to standard spectral clustering, robust path-based clustering results in a clearer distinction between small eigengaps and large ones and therefore corresponds to more stable solutions. Path-based clustering also more easily detects clusters with different number of members.

We find that a 2-cluster solution is preferred, with clusters that break fairly cleanly across the line $\mathcal{C} \approx 0.33$ (see Figure 13). Because this value of $\mathcal{C}$ corresponds to an architecture in which an evenly spaced ($\mathcal{C} = 0$) 4-planet system has been reduced to a 3-planet system by removing one of the middle planets, an immediate explanation for this subcluster presents itself: these are systems which are missing an intermediate planet. Either the planet has not yet been detected or it does not exist. However, because most of these systems also have low $\mathcal{Q}$, and hence do not host a (known) giant planet that might have frustrated planetesimal growth, we deem the non-detection hypothesis more likely. We explore this hypothesis in detail in section 6 below.

5. COMPARISON TO SYNTHETIC CATALOGUES

[PREAMBLE BLURB ABOUT FORWARD MODELING]

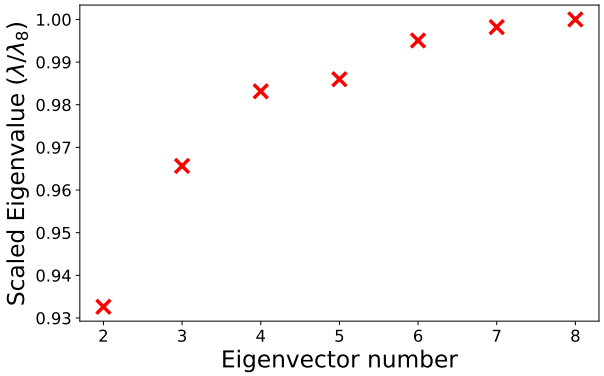

Figure 12. Eigenvalues for the first eight eigenvectors arising from our robust path-based clustering matrix . The first eigenvalue is near zero and for clarity is not shown on the plot. The largest subsequent eigengap is between the 2nd and 3rd eigenvectors, indicating that the optimal number of clusters for the path-based solution is two. The eigengap $\lambda_3 - \lambda_2$ is large compared to all other $\lambda_i - \lambda_{i-1}$, demonstrating that the clustering solution is stable.

[DEFEND WHY OUR COMPLEXITY MEASURES SHOULD AUTOMATICALLY MATCH IF ALL OBSERVATION BIASES HAVE BEEN ACCOUNTED FOR PROPERLY. THE ONLY UNDERLYING PHYSICS IS IN THE ESTIMATES OF THE MASSES AND MUTUAL HILL RADII.]

5.1. *EPOS*

We compare our real systems to a synthetic catalogue generated using the Exoplanet Population Observation Simulator (EPOS; Mulders et al. 2018, 2019) (Mulders et al. 2018, 2019). The catalogue was generated using multi-planet mode with all input parameters set to the optimized values found by Mulders et al. (2018), except for the radius ratio distribution, which was modified to draw from a a log-normal distribution with dispersion 0.15 dex. [BE EXPLICIT ABOUT HOW PERIOD AND INCLINATION ARE DRAWN]. All stellar hosts were identical to the Sun, i.e. $R_\star = R_\odot$, $M_\star = M_\odot$.

To ensure a valid domain for comparison, we restrict both the real catalogue and the synthetic catalogue to the radius and period limits used by Mulders et al. (2018) to optimize the EPOS fit parameters. We first eliminate any small planets ($r_p < 0.5 R_\oplus$) or long-period planets ($P > 400$ d), which is roughly analogous to setting more stringent detection thresholds. We then discard any systems which have been reduced to only one planet. Finally, we remove any systems which host at least one planet with $r_p > 6 R_\oplus$ or $P < 2$ d. The reason we eliminate entire systems rather than only removing the offending individual objects is because these large-radius or short-period planets are likely to be the most detectable and therefore correspond to artificial limitations of the simulator rather than natural observational bias.

We next recompute each of our complexity measures on the reduced collection of systems. Measures for the synthetic systems were computed following an identical procedure as for the real systems. In practice, this means that we we convert radii to masses following [NEIL & ROGERS] and we calculate flatness measure from transit durations following Equation 16.

We find that the CKS and EPOS populations are well matched for 3 out of 7 of our measures (see Figure 14).

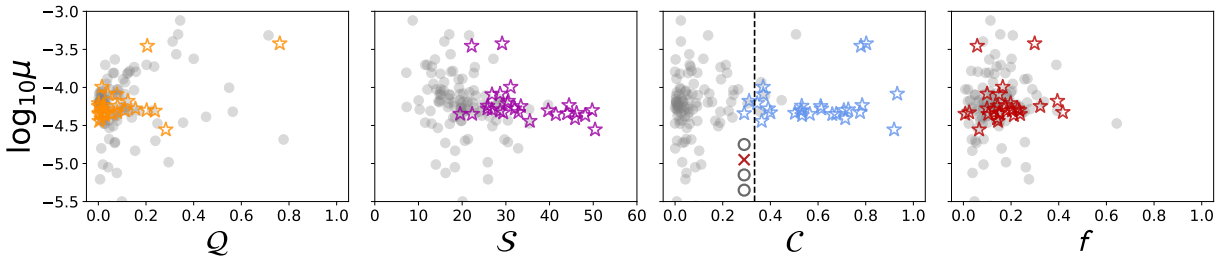

Figure 13. Results of robust path-based clustering assuming two clusters. The larger primary cluster (grey circles) centers around low $\mathcal{Q}$, $\mathcal{C}$, and f, with $\mathcal{S} \approx 20$. The smaller secondary cluster (red stars) extends out to high values of $\mathcal{C}$ and $\mathcal{S}$. We interpret this secondary cluster as a population of systems in which an intermediate planet has not been detected.

Based on AD and KS tests, the (observed) real and synthetic populations are drawn from the same underlying distributions of mass partitioning $\mathcal{Q}$, characteristic spacing $\mathcal{S}$, and multiplicity N. However, the EPOS systems are slightly - but significantly - more massive than the CKS systems, as traced by dynamical mass $\log \mu$. In addition, the real systems are more evenly spaced than predicted by EPOS, as traced by gap complexity $\mathcal{C}$. The evidence as to whether the real and synthetic populations are drawn from different monotonicity distributions is unclear, although we do note that the *Kepler* systems are slightly more concentrated towards $\mathcal{M} = 0$ than the EPOS systems. We find strong statistical evidence that the two populations are drawn from distinct flatness distributions, but because our measure of f depends on stellar properties, which were assumed to be uniform for EPOS, we hesitate to draw any definitive conclusions regarding the mutual inclinations of the systems.

[SUMMARIZE OUR INTERPRETATION OF CKS VS EPOS].

5.2. *SysSim*

We compare our real catalogue to a synthetic catalogue generated using the Exoplanet System Simulator (SysSim; Hsu et al. 2018, 2019; He et al. 2019). The catalogue was generated with input parameters set to the optimized values found by He et al. (2019). [DESCRIBE IN SOME DETAIL]. Our procedure for reconciling the limits of the simulated catalogue and the real CKS catalogue is identical to the procedure described for EPOS in the preceding section, although the bounds on period ($3 < P/\text{days} < 300$) and radius ($0.5 < r_p/R_\oplus < 10$) are slightly different. The procedures used to convert radii to masses and transit durations to flatness measures are likewise identical to the procedures described above.

We find that the CKS and SysSim populations are well matched for $\mathcal{Q}$ and somewhat matched for $\mathcal{S}$, but show statistically significant differences for $\log \mu$, $\mathcal{C}$, f, and $\mathcal{M}$ (Figure 15). As with EPOS, SysSim over predicts total system dynamical masses and under predicts the degree to which systems are clustered near very low gap complexity ($\mathcal{C} \to 0$). SysSim also produces a population of systems which are flatter than the real observed systems. However, because accurate stellar modeling was not a priority in the latest version of SysSim, we hesitate to over interpret these results. Given that both EPOS and SysSim fail to fully reproduce the observed distribution of f - the parameter which is most closely tied to stellar properties - we suggest that incorporating more rigorous constraints on stellar masses and radii into future population synthesis models may be a fruitful avenue for exploration.

Curiously, over half of SysSim systems show negative monotonicity, in stark contrast to the real data from CKS and to the synthetic data from EPOS. A major difference between EPOS and SysSim is that EPOS applies a cutoff to planets with $R > 6R_\oplus$, whereas SysSim simulates planets up to $R = 10R_\oplus$. However, these giant planets do not appear to be driving the large spike near $\mathcal{M} \approx -0.25$ for the SysSim population. We also recalculated directly using planetary radii without converting to mass first, but this did not alleviate the problem. Another key difference between SysSim and EPOS is that SysSim allows for multiple clusters of planets in a system whereas EPOS explicitly assumes only one cluster per system. We hypothesize that the presence of multiple clusters may be driving the observed negative monotonicities.

5.3. *Direct downsampling*

In addition to comparing against forward models, we also compare the data against itself by directly downsampling the highest multiplicity systems ($N \geq 4$) to generate a synthetic population of 3-planet systems. To produce this population, we randomly draw - with replacement - one of the 4+ planet systems from the CKS catalogue. We then generate a random number between 0 and 1 for each of the planets in the system, discarding any individual planets whose geometric transit probability ($P_{transit} \sim 1/a$) is less than the random number. If the downsampled system contains exactly 3 planets, we add it to our downsampled catalogue and repeat this procedure until the number of systems in the downsampled catalogue matches the number of 3-planet systems in our CKS catalogue. This procedure assumes moderate-to-high mutual inclinations between the planets and is admittedly less sophisticated than the routines run by EPOS and SysSim. Nevertheless, directly comparing the data against itself allows us to investigate whether the 3 planet and 4+ planet systems are drawn from the same intrinsic physical distribution.

The results of our downsample comparison are shown in Figure 16. We find that the downsampled 3 planet systems and real e planet systems are closely matched in $\mathcal{Q}$, $\mathcal{S}$, and $\mathcal{C}$ but that the downsampled systems are significantly flatter than the real systems. Notably, direct downsampling reproduces the distribution of $\mathcal{C}$ values much better than either of the population synthesis models. We interpret these results as evidence that all of the systems in our sample - regardless of multiplicity - are drawn from the same underlying physical distribution and that inter-system differences based on multiplicity can largely be explained by detection selection effects.

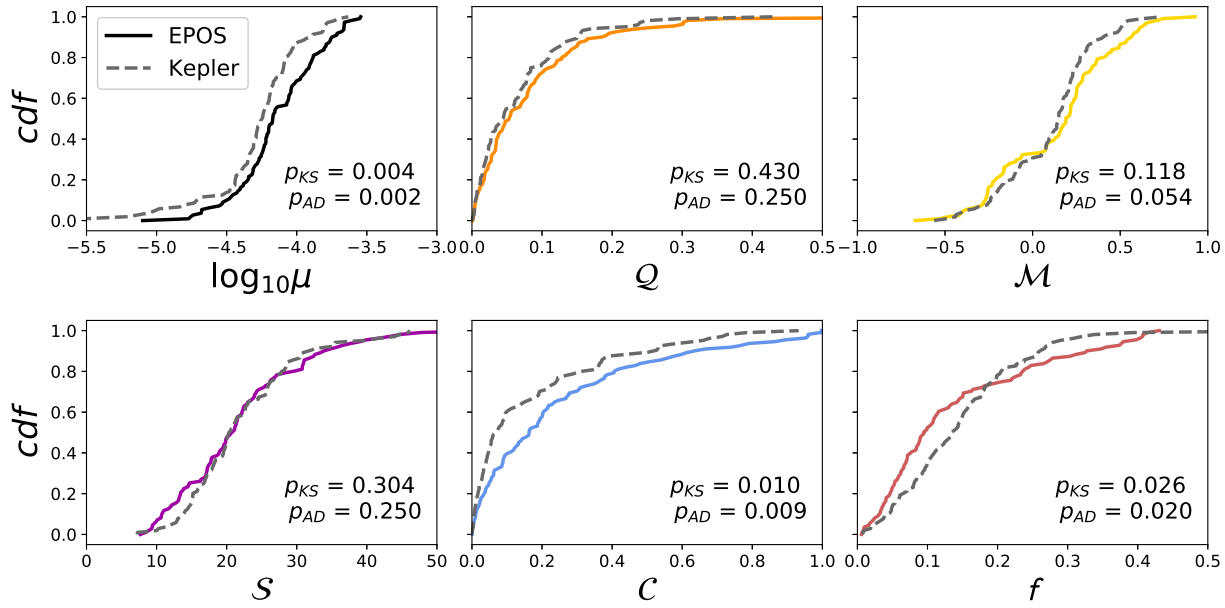

Figure 14. Cumulative density functions of system-level complexity measures for *Kepler* compared to EPOS. Dashed grey lines give results for *Kepler* and solid colored lines give results for EPOS. Resultant p-values of Anderson-Darling and Kolmogorov-Smirnov tests are shown on the plots. The data are well matched for $\mathcal{Q}$ and $\mathcal{S}$ but show statistically significant differences for $\log\mu$, $\mathcal{C}$, and f. It is unclear whether the underlying monotonicity distributions are different for the real vs. synthetic catalogues. Note that each panel has a different scale on the horizontal axis.

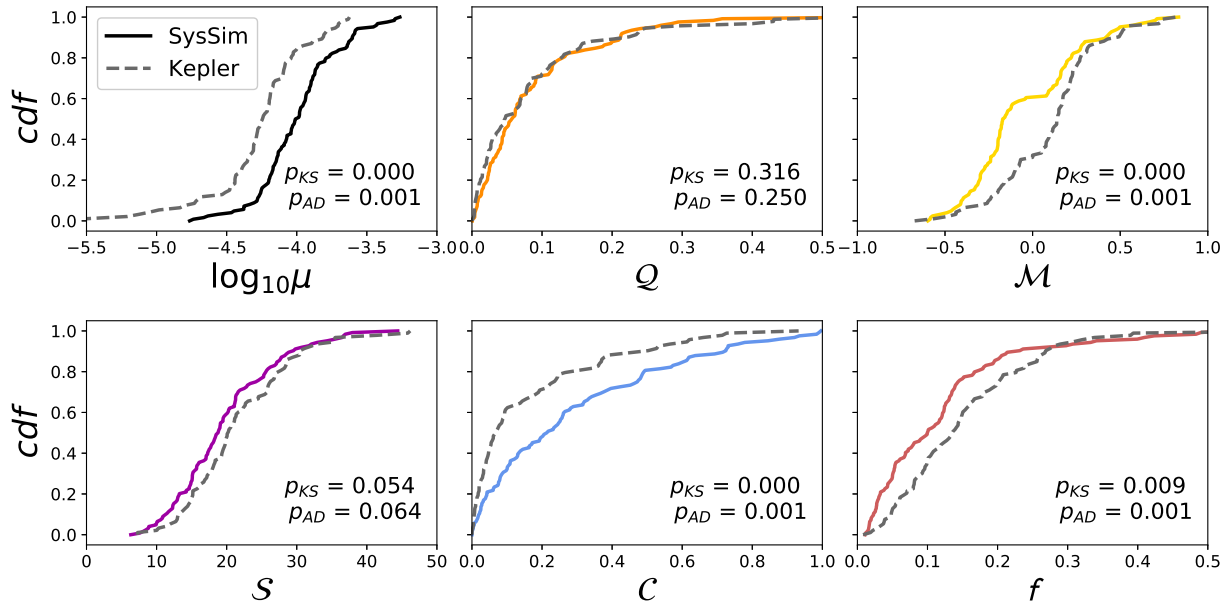

Figure 15. Cumulative density functions of system-level complexity measures for *Kepler* compared to SysSim. Dashed grey lines give results for *Kepler* and solid colored lines give results for SysSim. Resultant p-values of Anderson-Darling and Kolmogorov-Smirnov tests are shown on the plots. The data are well matched for $\mathcal{Q}$ and somewhat matched for $\mathcal{S}$ but show statistically significant differences for $\log\mu$, $\mathcal{C}$, f, and $\mathcal{M}$. Note that each panel has a different scale on the horizontal axis.

[THIS PARAGRAPH IS SLOPPY - RETHINK AND REWRITE]. Because none of our three methods for generating synthetic populations reproduces the observed distribution of f, we argue that a better understanding of mutual inclinations will be the key to understanding exoplanetary system architectures. We offer several explanations for why the real distribution of f is poorly matched with synthetic distributions. First, it is possible that f is not a very good measure of flatness. Second, maybe a Rayleigh distribution is not the appropriate function to model inclinations. Third, perhaps we do not understand the covariance in inclination between planets. In-depth investigations of planetary inclinations should be a priority of future work, and as such we are undertaking a project to remeasure impact parameters for the *Kepler* planets and thus provide a more direct way to assess inclination trends.

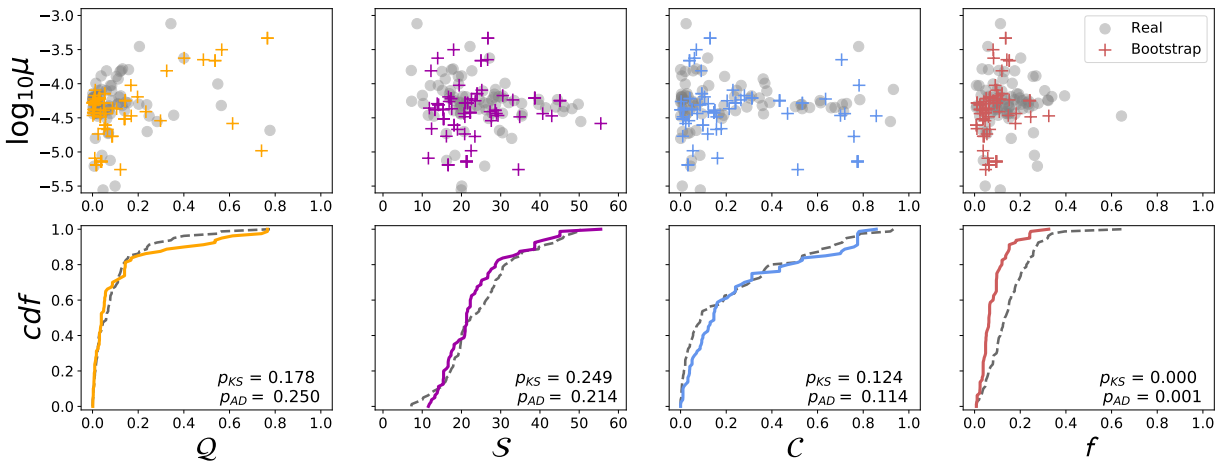

Figure 16. Downsampled bootstrap 3-planet systems. This is for a single iteration, but the results are representative of a typical outcome

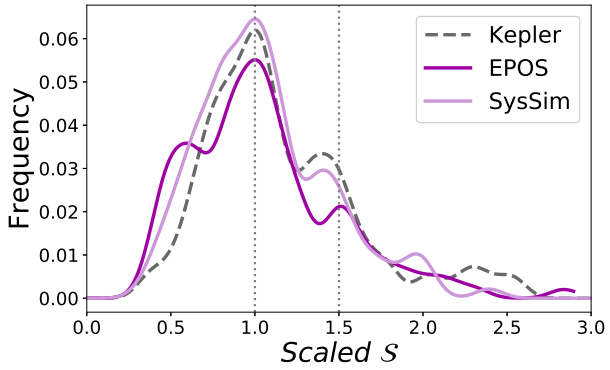

Figure 17. Echo peaks

6. CAN MISSING PLANETS EXPLAIN THE OBSERVED SYSTEM-LEVEL TRENDS?

[DISCUSS THE EVIDENCE FOR A SUBPOPULATION OF MISSING PLANETS] (1) Clusering finds subcluster at high C and S (1) Echo peak in S (2) mu is matched after normalizing by multiplicity (3) Q is matched pretty much all the time

Our explanation for the high-$\mathcal{C}$ subcluster makes the straightforward, testable prediction that these systems should host additional planets in their gaps. The evidence for such planets may already be present in the data, either as low signal-to-noise grazing transits or as dynamical signatures on the known planets that may be resolved via transit timing variations. In either case, because we now know where to look, it may be possible to identify marginal signals that would not pass muster in a blind search, but which - when placed in the context of their host systems - rise to the level of statistical significance. A search for these missing planets is beyond the scope of this project, and we leave this search for future work.

7. SUMMARY AND OUTLOOK

We have proposed seven dimensions along which to characterize exoplanetary systems: multiplicity, N; dynamical mass, μ; mass partitioning, $\mathcal{Q}$; monotonicity, $\mathcal{M}$; characteristic spacing, $\mathcal{S}$; gap complexity, $\mathcal{C}$; and flatness, f. [REFLECT ON THESE A BIT].

[SUMMARIZE OUR KEY FINDINGS]. (1) Planets really are like peas in a pod (2) There is an even greater degree of intra-system spacing uniformity than suggested by pair statistics alone (3) The jury is still out on monotonic ordering (4) It looks like all high-multiplicity *Kepler* systems belong to the same intrinsic population, with a subpopulation (∼[XX%]) differing along the spacing dimensions ($\mathcal{C}$ and $\mathcal{S}$) in a manner that can be explained by missing intermediate parameters.

One important parameter has been conspicuously absent from our above discussion: eccentricity. A full set of system-level measures must include a term to capture the degree of eccentricity of a system. A candidate is the angular momentum deficit (AMD; Laskar 1997; Laskar & Petit 2017) , defined as the amount of angular momentum the would need to be added to a system in order to circularize all of the planets. At the moment, high-precision eccentricity measurements are few and far between. However, our current efforts to refit impact parameters will also yield improved fits for eccentricities. It may also be that the best measure for describing exoplanet systems is not to use eccentricity and inclination independently, but rather to combine them into a sort of "dynamical hotness" term.

Many other measures of system-level architecture are possible. One possible candidate term is the typical distance from resonance for planets in a system. Such a resonance term could be expressed either as an average distance from resonance, or as a complexity-like term than answers the question "if one planet in a system is near resonance, what is the likelihood that all other planets in that system are also near resonance?" Other more exotic descriptors (each of which would require new measurements which are presently unobtainable) might be complexity terms based on alignments of nodes or periapses, varieties of bulk densities, atmospheric mass fractions, or even observed chemistry.

Future work should focus on a few things. (1) Applying this scheme to other populations of systems (e.g. RV) (2) Incoporating these ideas into population synthesis and planet formation models (3) Obtaining better measurements of planetary inclinations. (4) Searching for missing planets in the systems which we have identified as likely to be hosting additional undetected planets

[CONCLUDING THOUGHTS]

We thank Fred Ciesla, Leslie Rogers, Rebecca Willett, Andrey Kravtsov, Matthias He, and Eric Ford for helpful discussions about our methodology. We thank Gijs Mulders for providing synthetic EPOS catalogues, and we thank Andrew Neil for providing mass estimates for the planets.

Software: scipy, astropy, scikit-learn, KDEpy

REFERENCES

Alibert, Y. 2019, A&A, 624, A45
Anteneodo, C., & Plastino, A. 1996, Physics Letters A, 223, 348
Canup, R. M., & Ward, W. R. 2006, Nature, 441, 834
Chaitin, G. J. 1966, J. ACM, 13, 547
Chang, H., & Yeung, D.-Y. 2008, Pattern Recogn., 41, 191
Chen, J., & Kipping, D. 2017, ApJ, 834, 17
Chiang, E., & Laughlin, G. 2013, MNRAS, 431, 3444
Ciardi, D. R., Fabrycky, D. C., Ford, E. B., et al. 2013, ApJ, 763, 41
Dawson, R. I., Lee, E. J., & Chiang, E. 2016, ApJ, 822, 54
Fabrycky, D. C., Lissauer, J. J., Ragozzine, D., et al. 2014, ApJ, 790, 146
Fang, J., & Margot, J.-L. 2012, ApJ, 761, 92
Fulton, B. J., & Petigura, E. A. 2018, AJ, 156, 264
Furlan, E., Ciardi, D. R., Everett, M. E., et al. 2017, AJ, 153, 71
Gaia Collaboration, Prusti, T., de Bruijne, J. H. J., et al. 2016, A&A, 595, A1
He, M. Y., Ford, E. B., & Ragozzine, D. 2019, arXiv e-prints, arXiv:1907.07773
Hsu, D. C., Ford, E. B., Ragozzine, D., & Ashby, K. 2019, AJ, 158, 109
Hsu, D. C., Ford, E. B., Ragozzine, D., & Morehead, R. C. 2018, AJ, 155, 205
Hubble, E. P. 1926, ApJ, 64, 321
—. 1936, Realm of the Nebulae
Johnson, J. A., Petigura, E. A., Fulton, B. J., et al. 2017, AJ, 154, 108
Jontof-Hutter, D., Ford, E. B., Rowe, J. F., et al. 2016, ApJ, 820, 39
Kipping, D. 2018, MNRAS, 473, 784
Kolmogorov, A. N. 1968, International Journal of Computer Mathematics, 2, 157
Lansing, J. S. 2003, Annual Review of Anthropology, 32, 183
Laskar, J. 1997, A&A, 317, L75
Laskar, J., & Petit, A. C. 2017, A&A, 605, A72
Lissauer, J. J., Ragozzine, D., Fabrycky, D. C., et al. 2011, ApJS, 197, 8
López-Ruiz, R., Mancini, H., & Calbet, X. 1995, Physics Letters A, 209, 321
Lopez-Ruiz, R., Mancini, H., & Calbet, X. 2010, arXiv e-prints, arXiv:1009.1498
Manson, S. M. 2001, Geoforum, 32, 405
May, R. M. 1976, Nature, 261, 459
Miguel, Y., Cridland, A., Ormel, C. W., Fortney, J. J., & Ida, S. 2019, arXiv e-prints, arXiv:1909.12320
Millholland, S., Wang, S., & Laughlin, G. 2017, ApJ, 849, L33
Mosqueira, I., & Estrada, P. R. 2003, Icarus, 163, 198
Mulders, G. D., Mordasini, C., Pascucci, I., et al. 2019, arXiv e-prints, arXiv:1905.08804
Mulders, G. D., Pascucci, I., Apai, D., & Ciesla, F. J. 2018, AJ, 156, 24
Ng, A. Y., Jordan, M. I., & Weiss, Y. 2001, in Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, NIPS'01 (Cambridge, MA, USA: MIT Press), 849–856
Panos, C. P., Chatzisavvas, K. C., Moustakidis, C. C., & Kyrkou, E. G. 2007, Physics Letters A, 363, 78
Petigura, E. A., Howard, A. W., Marcy, G. W., et al. 2017, AJ, 154, 107
Pu, B., & Wu, Y. 2015, ApJ, 807, 44
Shannon, C. E. 1948, The Bell System Technical Journal, 27, 379
Shiner, J. S., Davison, M., & Landsberg, P. T. 1999, Phys. Rev. E, 59, 1459
Szkely, G. J., Rizzo, M. L., & Bakirov, N. K. 2007, Ann. Statist., 35, 2769
Thompson, S. E., Coughlin, J. L., Hoffman, K., et al. 2018, ApJS, 235, 38
von Luxburg, U. 2007, Statistics and Computing, 17, 395
Wang, S. 2017, Research Notes of the American Astronomical Society, 1, 26
Weiss, L. M., Marcy, G. W., Petigura, E. A., et al. 2018, AJ, 155, 48
Winn, J. N., & Fabrycky, D. C. 2015, ARA&A, 53, 409
Zhu, W. 2019, arXiv e-prints, arXiv:1907.02074