Evidence for Two Hot-Jupiter Formation Paths

Benjamin E. Nelson1,2, Eric B. Ford3,4, and Frederic A. Rasio1

1 Center for Interdisciplinary Exploration and Research in Astrophysics (CIERA) and Department of Physics and Astronomy, Northwestern University, 2145 Sheridan Road, Evanston, IL 60208, USA; benelson@northwestern.edu
2 Northwestern Institute for Complex Systems, 600 Foster Street, Evanston, IL 60208, USA
3 Center for Exoplanets and Habitable Worlds, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA, 16802, USA
4 Department of Astronomy & Astrophysics, The Pennsylvania State University, 525 Davey Laboratory, University Park, PA 16802, USA

Received 2017 January 31; revised 2017 June 26; accepted 2017 July 5; published 2017 August 23

Abstract

Disk migration and high-eccentricity migration are two well-studied theories to explain the formation of hot Jupiters. The former predicts that these planets can migrate up until the planet–star Roche separation (aRoche) and the latter predicts they will tidally circularize at a minimum distance of 2 aRoche. Considering long-running radial velocity and transit surveys have identified a couple hundred hot Jupiters to date, we can revisit the classic question of hot-Jupiter formation in a data-driven manner. We approach this problem using data from several exoplanet surveys (radial velocity, Kepler, HAT, and WASP) allowing for either a single population or a mixture of populations associated with these formation channels, and applying a hierarchical Bayesian mixture model of truncated power laws of the form $x^{-\gamma}$ to constrain the population-level parameters of interest (e.g., location of inner edges, $\gamma$, mixture fractions). Within the limitations of our chosen models, we find that the current radial velocity and Kepler sample of hot Jupiters can be well explained with a single truncated power-law distribution with a lower cutoff near 2 aRoche, a result that still holds after a decade, and $\gamma = -0.51 \pm 0.06$. However, the HAT and WASP data show evidence for multiple populations (Bayes factor $\approx 10^{21}$). We find that $15 \pm 6\%$ reside in a component consistent with disk migration ($\gamma = -0.04 \pm 0.03$) and $85 \pm 6\%$ in one consistent with high-eccentricity migration ($\gamma = -1.38 \pm 0.23$). We find no immediately strong connections with some observed host star properties and speculate on how future exoplanet surveys could improve upon hot-Jupiter population inference.

Key words: methods: statistical – planetary systems – planets and satellites: formation – planets and satellites: gaseous planets

1. Introduction

More than two decades after their discovery, hot Jupiters remain elusive in regard to their primary formation channel. They likely formed at several astronomical units, beyond their host stars’ ice lines (Mizuno 1980; Pollack et al. 1996). Theories for bringing these giants into sub-Mercury orbits can generally be lumped into two categories. The first is a gentle, slow orbital decay through interactions with a protoplanetary disk (Goldreich & Tremaine 1980; Lin et al. 1996; Ward 1997; Murray et al. 1998). The second is tidal circularization of a planet on a highly eccentric orbit, created from sudden or secular gravitational encounters with other massive bodies (Rasio & Ford 1996; Wu & Murray 2003; Fabrycky & Tremaine 2007; Naoz et al. 2011; Wu & Lithwick 2011).

Very simple theories of disk versus eccentric migration also make testable predictions. Previous studies have investigated these theories in several contexts, including but not limited to spin–orbit misalignment (Naoz et al. 2012; Rogers et al. 2012), stellar metallicity (Dawson & Murray-Clay 2013), and the presence of additional companions (Knutson et al. 2014; Ngo et al. 2015, 2016; Piskorz et al. 2015; Bryan et al. 2016; Schlafmann & Winn 2016). The focus of this paper will be in the context of the Roche separation aRoche between the planet and host star. For an infinitely compressible planet, this is given by $a_{\text{Roche}} = \frac{2.16}{R}M_{\text{p}}^{-1/3}$, where $R$ is the physical radius of the planet and $\mu$ is the planet–star mass ratio (Faber et al. 2005).

In disk migration, planets can inwardly migrate up until $a_{\text{Roche}}$. A Jupiter-sized planet may be consequently stripped of its gaseous envelope or tidally disrupted depending on the details of mass transfer at this distance (Valsecchi et al. 2014, 2015). In eccentric migration, $a_{\text{Roche}}$ is the minimum periastron distance for the planet to survive without considerable mass loss. Such an orbit would circularize at a semimajor axis greater than 2$a_{\text{Roche}}$. Pioneering radial velocity surveys showed the observed distribution of hot-Jupiter semimajor axes normalized to their respective Roche separations ($x \equiv a/a_{\text{Roche}}$) has an inner cutoff near 2 (Ford & Rasio 2006, FR06 henceforth), supporting the prediction from an eccentric migration history. At the time, the sample size was 21 planets, most with only minimum mass ($M \sin i$) measurements.

One decade since then, there are many more radial velocity planets as well as dedicated photometric transit surveys that are able to probe physical radii. Stars targeted by ground-based transit surveys are often amenable to radial velocity follow-up, so a physical mass (rather than $M \sin i$) can be inferred for these planets, providing individual $x$ values.

By combining planet discoveries from various surveys into an updated sample, we may be able to say something more precise about the population statistics of hot Jupiters with respect to $x_{\text{Roche}}$ and how they compare to predictions of well-studied formation theories. Rather than motivating and testing a particular theory to compare to the observations (e.g., through planet population synthesis models), we approach this problem from a data-driven perspective. In particular, could we infer a single population or disentangle two populations from the data, where each population is consistent with a disk migration and/or eccentric migration history?

In this paper, we show how a hierarchical Bayesian framework can address this classic question of hot-Jupiter formation. In Section 2, we describe the data sets used for the study. In
Table 1

| Parameter | Description |
|-----------|-------------|
| \( M_1 \) | Models |
| \( M_t \) | One-component truncated Pareto |
| \( M_t,\sigma \) | \( M_t \) with observational uncertainties |
| \( M_2 \) | Two-component truncated Pareto mixture |
| \( M_2,\sigma \) | \( M_2 \) with observational uncertainties |

\( d \) Data

\( \hat{P} \) Observed orbital period

\( \hat{M} \) Observed planet mass

\( \hat{R} \) Observed physical radius

\( \sigma_P \) Uncertainty in observed orbital period

\( \sigma_M \) Uncertainty in observed planet mass

\( \sigma_R \) Uncertainty in observed physical radius

\( \alpha \) Individual Parameters

\( i \) True inclination of orbital plane to sky plane

\( P \) True orbital period

\( M \) True planet mass

\( R \) True physical radius

\( x \) Planet semimajor axis divided by Roche separation

\( x_{\text{edge}} \) \( x \) assuming true mass = minimum mass

\( \beta \) Hyperparameters

\( x_{1} \) Lower limit of \( x \) for one-component distribution

\( x_{1,1} \) Lower limit of \( x \) for first majorixture component

\( x_{1,2} \) Lower limit of \( x \) for second majorixture component

\( x_s \) Upper limit of \( x \)

\( \gamma \) Power-law index for one-component distribution in \( x \)

\( \gamma_1 \) Power-law index for first majorixture component in \( x \)

\( \gamma_2 \) Power-law index for second majorixture component in \( x \)

\( \eta_i \) ith mixture component fraction

Section 3, we describe the statistical models applied to each data set. In Section 4, we show the results of our inference and discuss the details and caveats in Section 5. Table 1 provides descriptions of parameters commonly used throughout the manuscript.

2. Data

For this analysis, we consider four different samples of exoplanets discovered from various surveys: radial velocity, HATNet (Bakos et al. 2004), WASP (Street et al. 2003), and Kepler (Santerne et al. 2016). For all of these data sets, we restrict the orbital periods to be \( P < 30 \) d and velocity semi-amplitude \( K > 30 \) m s\(^{-1}\) to minimize selection effects (Cumming 2004, FR06). We recognize that these transit surveys are prone to observational biases and selection effects not present in long-running RV surveys, which we will address later in this section.

There are multiple well-known exoplanet databases to choose from and each have their own strengths. The NASA Exoplanet Archive (Akeson et al. 2013) consistently keeps up with new planet discoveries and refinements on orbital elements but is therefore likely to contain some controversial planet candidates, exoplanets.org (Han et al. 2014) is slower to update planet discoveries and properties but is much more thorough in vetting planet claims. The choice here comes down to an incomplete and slightly out-of-date list of planets versus an uncurated but up-to-date list. We ultimately analyze planets queried from both databases but present the Exoplanet Archive results in this manuscript. We will describe any difference data sets and results in Sections 5.1 and 5.3 respectively.

We queried the Exoplanet Archive on March 4, 2017 to obtain the orbital properties for these planets given the above period and semi-amplitude restrictions above. For the radial velocity sample, we only include planets discovered through the radial velocity method, stressing that this set does not include planets discovered by transit surveys that had follow-up radial velocity measurements. For the HAT and WASP samples, we filter for any planet with a WASP or HAT substring in its name and ensure that they were discovered via the transit method. The only exception was WASP-94, a wide binary system with both stars hosting their own planets (Neveu-VanMalle et al. 2014). The hot Jupiter around the secondary star was detected through radial velocity and is thus excluded from our sample.

For those planets without measured radii, we assume a radius of \( 1.2 \pm 0.1R_J \), which is roughly how known exo-Jupiter radii are distributed. For the Kepler sample, we consult the planets with “secured” planetary status listed in Table 7 of Santerne et al. (2016) and apply the same period and semi-amplitude cutoffs. No planet was seen in multiple data sets.

Now \( x \) can be rewritten in terms of radius \( R \), orbital period \( P \), and planet mass \( M \) as such:

\[
x = \frac{a}{a_{\text{Roche}}} = 0.462 \left( \frac{P}{\text{1 year}} \right)^{2/3} \left( \frac{M}{1 M_\odot} \right)^{1/3} \left( \frac{R}{1 \text{ au}} \right)^{-1}.
\]

We will consider two models, one that neglects and one that incorporates observational uncertainties in these parameters (see Section 3).

We ultimately obtain 52 planets for our RV data, 77 for HAT, 74 for WASP, and 27 for Kepler. Figure 1 shows how \( x \) is distributed for each data set, assuming edge-on orbits (i.e., the true mass is the same as the measured minimum mass). At this point, we simply want a rough idea of how these data are distributed, so we visualize them in two different ways: a cumulative distribution to emphasize the location of the smallest \( x \) and a more experimental network-based frequency analysis (NFA) to emphasize rough functional form. The cumulative distribution plots a running probability over \( x_{\text{edge}} \), i.e., \( x \) assuming an edge-on orbit. The NFA considers each \( x_{\text{edge}} \) and if they are “connected” with other data, rather than counting up \( x_{\text{edge}} \) in a bin (e.g., as in a traditional histogram), and plots the number of network connections against each value of \( x_{\text{edge}} \). Two values are connected if they fall within a given range, \( \pm \epsilon \), in \( x \)-space. The output therefore depends on the chosen value of \( \epsilon \), but this can be selected in an automated manner by verifying that properties of the network have become stable. Details of these methods are discussed in Derrible & Ahmad (2015). The “optimal” values of \( \epsilon \) for these data sets are 1.81 for RV, 0.46 for HAT, 0.84 for WASP, and 0.94 for Kepler.

The cumulative distributions shows \( x_{\text{edge}} \) truncating around 2 for the RV and Kepler samples, and at 1 < \( x_{\text{edge}} < 2 \) for the HAT and WASP samples. The NFA agrees with this but also shows that the HAT and WASP distributions have a peak somewhere between 2 < \( x_{\text{edge}} < 3 \). The peak seen in the RV

\(^{5}\) The code for performing NFA is also publicly available at https://github.com/cs2164/INF.\footnote{The code for performing NFA is also publicly available at https://github.com/cs2164/INF.}
can be explicitly grouped into two subsets, one
a

Figure 1. Distribution in $x$ assuming edge-on orbits (i.e., $x_{\text{edge}}$) for planets discovered through the following surveys: RV (blue dots), HAT (green solid line), WASP (red dot-dashed line), and Kepler (cyan dashed line). Left: the cumulative distributions across the full range of $x_{\text{edge}}$ (top) and zoomed into the $0 < x_{\text{edge}} < 6$ range (bottom). Right: the distribution visualized through a network-based frequency analysis across the full range of $x_{\text{edge}}$ (top) and zoomed into the $0 < x_{\text{edge}} < 6$ range (bottom). The effective widths $\zeta$ used here are 1.81 (RV), 0.46 (HAT), 0.84 (WASP), and 0.94 (Kepler). For more on the method, refer to Derrible & Ahmad (2015).

demographics may differ from those of RV surveys (Wang et al. 2015, and Figure 2).

We can easily account for the geometric bias for planets discovered from transit surveys, since $x$ scales with $a$. The other aforementioned biases depend on details of each survey’s observing strategy. However, the first goal of this study is to not necessarily obtain an accurate statistical description of the true hot-Jupiter population(s) but rather to see if the current sample could allow for testing specific models of planet formation. The observing baseline biases against longer-period planets would mostly affect how the distribution of hot Jupiters tapers off at higher values of $x$. Fortunately, the HAT and WASP surveys now have sufficient temporal and longitudinal coverage to be sensitive to hot Jupiters over a wide range of orbital periods reaching 30 days.

We can look at a few properties of the host stars sampled from these surveys to verify differences in host star demographics. Figure 2 looks at correlations in stellar effective temperature, metallicity, and rotation speed with $x_{\text{edge}}$ for each survey sample. We see that RV and Kepler surveys typically probe more metal-rich stars than the HAT and WASP surveys. We also find that HAT and WASP planets in the range of $1 < x_{\text{edge}} < 2$ are confined to a narrower region of $T_{\text{eff}}$ relative to the overall sample. We will compare the results of our statistical analysis with these host star properties in Section 5.

### 3. Statistical Model

We approach this problem through a hierarchical Bayesian modeling (HBM) framework. In HBM, a set of model parameters $\theta$ can be explicitly grouped into two subsets, one for parameters describing the individual objects (i.e., individual parameters, $\alpha$) and one describing the properties of the objects’ population(s) (i.e., population parameters or hyperparameters, $\beta$). Conditioned on the observed data ($d$), Bayes theorem is rewritten as

$$p(\alpha, \beta | d) = \frac{p(d | \alpha, \beta) p(\alpha) p(\beta)}{p(d)}$$  \hspace{1cm} (2)$$

where $p(d | \alpha, \beta)$ is the likelihood function, $p(d)$ is the fully marginalized likelihood, and $p(\alpha, \beta | d)$ is the joint posterior probability distribution. Using the law of conditional probability, $p(\alpha | \beta) p(\beta)$ replaces the prior probability distribution of $p(\theta) = p(\alpha, \beta)$, where $p(\alpha | \beta)$ is the prior on the individual parameters conditioned on the hyperparameters and $p(\beta)$ is the prior on the hyperparameters.

Explicitly grouping the model parameters in this manner allows us to see the relationship between the individual parameters and hyperparameters. The distribution of $x$ for each planet is a function of both the individual parameters (e.g., mass and orbital period) and the hyperparameters describing the population. Likewise, the individual parameters as a group inform the hyperparameters. This allows for simultaneous inference on the properties of individual planets in our sample and properties of the sample in aggregate. Recent studies in the exoplanet literature have employed such multi-level modeling and laid the groundwork for hierarchical inference in exoplanet science. These include the occurrence rate of Earth-analogs (Foreman-mackey et al. 2014), composition distribution of sub-Neptunes (Wolfgang & Lopez 2015), a probabilistic mass–radius relationship for sub-Neptunes (Wolfgang et al. 2016), the transition radius between rocky and non-rocky planets.

---

Ground-based transit surveys can suffer from selection effects very different from those of RV surveys. There is a geometric bias toward shorter orbital periods (e.g., the geometric probability of transit scales as $a^{-1}$), transits can be missed due to the rotation of the Earth, three transits are often required to confirm a planet that biases against long-period planets (with respect to the observing baseline), and the stellar
In the following subsections, we describe two general statistical models that we apply to the data.

### 3.1. One-component Power Law, $\mathcal{M}_1$

Our first model $\mathcal{M}_1$ is inspired by the analysis conducted by FR06. We model the distribution of $x$ for hot Jupiters as a truncated power law, with power-law index $\gamma$ and lower and upper limits of $x$ being $x_l$ and $x_u$, respectively. This is the same as a Pareto distribution with a lower truncation $x_l$ and upper truncation $x_u$ (“truncated Pareto” henceforth; Beg 1981; Aban et al. 2006). Here, $\gamma$, $x_l$, and $x_u$ are the hyperparameters $\beta$.

For now, let us assume that all of the observed planets are orbiting edge-on, that all unreported radii are 1.2 times the measured value, and that all unreported inclinations are zero. If the planet sample was discovered through a transit survey and we wanted to account for the geometric bias, then we would apply a factor of $1/a$ (i.e., $\gamma - 1$ to $\gamma - 2$ in the exponent and $\gamma$ to $\gamma - 1$ in the normalization).

For most radial velocity planets, the orbital inclination and planet radius are not measured. For convenience, we sort the sample such that all inclinations are measured and all radii are reported.

In the following subsections, we describe two general statistical models that we apply to the data.
non-transiting planets contained in $x$. For now, unknown radii are still fixed to $1.2R_J$.

Our posterior probability distribution is

$$p(\gamma, x_i, x_u, l_m + 1, \ldots, l_n | x) \sim \gamma^n (x_u^\gamma - x_l^\gamma)^{-\gamma} \prod_{j=1}^{n} \left( \frac{\text{sd}_{\text{obs}, j}}{\sin i_j} \right)^{\gamma - 1} = \gamma^n (x_u^\gamma - x_l^\gamma)^{-\gamma} \prod_{j=1}^{m} x_j^{\gamma - 1} \prod_{j=m+1}^{n} \left( \frac{\text{sd}_{\text{obs}, j}}{\sin i_j} \right)^{\gamma - 1}. \quad (4)$$

In this hierarchical framework, Stan samples from $\alpha = \{l_{m+1}, \ldots, l_n\}$ and $\beta = \{\gamma, x_i, x_u\}$ simultaneously. To verify the robustness of this model, we can also incorporate uncertainties in period, mass, and radius into a model $M_{l,\sigma}$. Here, we assume each observed value ($\hat{P}, \hat{M}, \hat{R}$) is drawn from a Gaussian distribution with a mean true value ($P, M, R$) and standard deviation ($\sigma_{\hat{P}}, \sigma_{\hat{M}}, \sigma_{\hat{R}}$) given by the reported uncertainties. In other words,

$$\hat{P}_j \sim N(P_j, \sigma_{\hat{P}}^2)$$
$$\hat{M}_j \sim N(M_j, \sigma_{\hat{M}}^2)$$
$$\hat{R}_j \sim N(R_j, \sigma_{\hat{R}}^2), \quad (5)$$

where we assume $\sigma_{\hat{R}} = 0.1R_J$ for planets with no radius measurement. We do not include uncertainties in orbital inclination. Although posterior distributions obtained from RV or transit data sometimes reveal asymmetric uncertainties in the planetary parameters (Zakamska et al. 2011), we focus on period, mass, and radius of planets with high signal-to-noise, so Gaussianity is a reasonable assumption for this study. This appears to be the case for the vast majority of reported uncertainties in masses and orbital period. We do note a varying degree of asymmetry in some of the radius measurement uncertainties. In those cases, we simply average the upper and lower error bar as our effective $\sigma_{\hat{R}}$. Figure 3 shows $M_l$ and $M_{l,\sigma}$ in a probabilistic graphical form (neglecting differences in $i$ and $R$ for planets $1, \ldots, m$ and $m+1, \ldots, n$ for simplicity).

A truncated Pareto distribution is convenient here for a few reasons. First, it provides a direct comparison to the results of FR06, who employed the same distribution. Second, it is unclear how nature distributes hot Jupiters in $x$-space. A truncated Pareto is flexible enough to allow for monotonically increasing and decreasing (or uniform) frequency in $x$ over the limited range of our sample. Lastly, the lower truncation parameter $x_l$ is motivated as a well-understood theoretical stopping point for planets undergoing disk or eccentric migration at $x_l = 1$ and $x_l = 2$ respectively. However, we discuss some caveats in Section 5.
3.2. Two-component Power-law Mixture

It is not clear that a single-component Pareto would accurately describe our data sets based on Figure 1, specifically, the HAT and WASP sample. Furthermore, there may be enough planets in aggregate to meaningfully constrain multiple overlapping populations. In particular, what if the current sample of hot Jupiters was the result of a mixture of two processes: one component consistent with disk migration and the other with eccentric migration?

Here, we consider a model with a mixture of two truncated Pareto distributions, \( \mathcal{M}_2 \). Each component \( k \) is modeled with its own \( g_k, \gamma_k, x_{1,k}, \) and the fraction of planets residing in the \( k \)th component \( f_k \). \( f \) is a simplex demanding that \( \sum_k f_k = 1 \). Both components share a common \( x_u \). We therefore sample from the following posterior distribution:

\[
p(f, \gamma, x_{1,1}, x_{1,2}, x_u, i_{m+1}, \ldots, i_n | \mathbf{x}) \\
\sim \prod_{j=1}^{n} \left\{ f_1 \frac{\gamma_1}{(x_u^{\gamma_1} - x_{1,1}^{\gamma_1})} \frac{x_{\text{edge},j}}{\sin^{1/3} l_j} \right\}^{\gamma_1 - 1} \\
+ f_2 \frac{\gamma_2}{(x_u^{\gamma_2} - x_{1,2}^{\gamma_2})} \frac{x_{\text{edge},j}}{\sin^{1/3} l_j} \right\}^{\gamma_2 - 1}. \tag{6}
\]

Figure 4 shows the probabilistic graphical models and visualizes the hyperparameters for \( \mathcal{M}_2 \) and \( \mathcal{M}_{2,\sigma} \).

4. Results

We apply \( \mathcal{M}_1 \) and \( \mathcal{M}_2 \) to the RV+Kepler data and HAT+WASP data separately and sample from the posterior distribution using \textit{Stan}.

We assess the evidence for both models by computing the fully marginalized likelihood (i.e., Bayesian evidence) of the population-level parameters using an importance sampling algorithm (Nelson et al. 2016). For the RV+Kepler data, the Bayes factor \( p(d|\mathcal{M}_2)/p(d|\mathcal{M}_1) \) is \( \approx 0.1 \). For the HAT+WASP data, the Bayes factor is on the order of \( 10^{21} \). Given the choice of our statistical models, we conclude that the RV+Kepler data do not provide sufficient evidence for multiple populations, but the HAT+WASP data show decisive evidence for two populations.

For the RV+Kepler data, we run our MCMC with 10 chains for at least 300,000 generations each for \( \mathcal{M}_1 \) and \( \mathcal{M}_{1,\sigma} \). For the HAT+WASP data, we run our MCMC with 10 chains for at least 400,000 generations each \( \mathcal{M}_2 \) and \( \mathcal{M}_{2,\sigma} \). We throw out the first half of the sample as burn-in then thin the remainder to ultimately obtain 10,000 posterior samples.

For Figures 5 and 6, filled contours correspond to models that do not include uncertainties in measurements of the planet properties (i.e., \( \mathcal{M}_1 \) or \( \mathcal{M}_2 \)) and line contours correspond to
models that do include observational uncertainties (i.e., $M_{l,\sigma}$ and $M_{z,\sigma}$).

4.1. One-component Model: RV and Kepler Data Sets

The functional form of $p(\beta|x)$ for the RV and Kepler data look similar, ending abruptly at $x \approx 2$ and generally tapering off at larger $x$. Figure 5 shows the posterior distribution of $\gamma$ and $x_l$ for the one-component model applied to the these data sets individually and together.

In the RV sample, there are a couple planets with $x_{\text{edge}} \approx 2$, but this is assuming their minimum masses are close to their physical masses. By modeling unknown inclinations (thus unknown physical masses) for these planets and all the planets as drawn from a single population, we find a probable range of $x_l$ from 2.1 to 2.9. In the Kepler sample, all of the planets are transiting, so their RVs probe physical masses. $x_l$ is governed by the planet with the smallest $x$, which is slightly less than 2, and we see a truncated contour. In the RV+Kepler sample, the location of $x_l$ remains the same as the Kepler-only result but we obtain a more precise constraint on $\gamma$. We can verify the robustness of these results by modeling measurement uncertainties, and they indeed overlap very well with $\gamma = -0.51_{-0.42}^{+0.19}$. In each case, there is an anti-correlation between $x_l$ and $\gamma$, which we interpret as an artifact of our chosen population model: tuning $x_l$ to higher values for a fixed $x_u$ forces the distribution to become much steeper, i.e., a $\gamma$ that is more negative.

The above result appears to be consistent with a single population that underwent an eccentric migration history. However, we note that the RV sample includes several hot Jupiters in multi-planet systems. Some even have multiple planets with semimajor axes less than 1 au (e.g., 55 Cancri, HIP 14810). We perform a separate run on an RV data set without known multi-planet systems and find quantitatively consistent results in regards to $x_l$ and $\gamma$.

We additionally test $M_2$ for the combined data set and find the population-level parameters to be mostly unconstrained. This is reflected in our Bayes factor, which shows a lack of strong evidence for $M_2$.

After a decade since FR06, it appears that the inner edge at $x_l \approx 2$ still holds for RV-discovered planets, and maybe even Kepler planets. We will address possible caveats regarding this in Section 5.

4.2. Two-component Mixture Model: HAT/WASP and All Data Sets

Figure 6 shows the posterior distribution of several parameters that specify our two-component mixture model applied to the HAT+WASP data. Our MCMC was initialized around $x_{l,1} \sim 1$ and $x_{l,2} \sim 2$, motivated from the predictions made by disk and eccentric migration respectively. Two-component properties of the HAT and WASP samples could be meaningfully constrained, as shown in the top and middle panels of both figures. The results for the HAT+WASP samples are shown in the bottom panels.

The left figure shows constraints on the component fractions $f$ and power-law indices $\gamma$ of each modeled component (blue and green contours). Neither component has $f$ consistent with 0 or 1, suggesting that our model can successfully separate two overlapping populations of planets. For the model including uncertainties applied to the HAT+WASP data, estimates for component fractions based on 15.9, 50, and 84.1 percentiles are
The joint posterior distributions of a two-component model (see Figure 4) for a data set containing solely WASP planets (top panels) and HAT+WASP planets (bottom panels). In sets of panels, filled contours show 68% (darker) and 95% (lighter) credible intervals without considering uncertainties in observed planet properties. Line contours show 68% (dashed) and 95% (solid) credible intervals with uncertainties in observed planet properties. Top: joint posterior distributions of power-law index $\gamma$ (horizontal axis) and component fraction $f$ (vertical axis) for the first (e.g., $x_{1}$) and second (e.g., $x_{2}$) components. Bottom: joint posterior distribution of $x_{1}$ for the first (horizontal axis) and second (vertical axis) components compared to the theoretical truncation locations of the two migration models (dashed lines).

0.15$^{+0.06}_{-0.09}$ and 0.85$^{+0.06}_{-0.09}$ for the first and second components respectively. The first component suggests a relatively shallow Pareto distribution over $x$ ($\gamma = -0.04^{+0.53}_{-1.27}$) and the second component suggests a much steeper distribution toward smaller $x$ ($\gamma_{2} = -1.38^{+0.22}_{-0.47}$).

The right figure shows constraints on each $x_{i}$ compared to the theoretical truncation locations of the two migration models (dashed lines). The posterior distribution of $x_{1,1}$ is unsurprisingly governed by the planet with the smallest $x$, but $x_{1,2}$ shows multi-modal structure around $x_{1,2} \approx 2$ (filled contours). The distribution looks smoother once observational uncertainties are considered (line contours). Some Markov chains were found to be sampling more confined regions of $x_{1,2}$ space and we suspect that these were local minima. We simply removed these chains from our posterior sample.

We can also quantitatively assess convergence of the hyperparameters using the Gelman–Rubin statistic. In practice, a value close to 1 implies the estimand has converged. We compute the following values for each hyperparameter: 1.035 (first component), 1.046 ($\gamma_{1}$), 1.038 ($\gamma_{2}$), 1.004 ($x_{1,1}$), 1.085 ($x_{1,2}$), and 1.020 ($x_{n}$). We provide our interpretations in Section 5.3.

4.3. Two-component Model: Synthetic Data Set

Could the multi-modality issues in the previous section be attributed to a relatively small or incomplete sample of hot Jupiters? To test this, we attempt to recover the input properties of three synthetically generated data sets using a two-component mixture model (without uncertainties). These data sets contain a set of planets with known inclinations drawn from a mixture of two truncated Paretos with the following properties: $\{f_{1, f_{2}} = 0.25, 0.75; \{\gamma_{1}, \gamma_{2}\} = 1, -1; \{x_{1,1}, x_{1,2}\} = 1, 2; \}$ and $x_{n} = 20$. The sample size is 100, 500, and 1000 respectively, and our MCMC is initialized close to the true values of the generative model.

Figure 7 show constraints on the same hyperparameters discussed in Section 4.2 for our three data sets in comparison to the true input values denoted by red diamonds. In each case, our constraints were consistent with the true values of the generative model. As was shown in previous figures, the distribution of $x_{1,1}$ is truncated at the value of the smallest $x$ in
each data set. Some multi-modal structure arises in $x_{1.2}$ for the data set with only 100 planets, somewhat similar to what was seen in Figures 6. This structure disappears when the sample size increases. So it is plausible that the multi-modal structure seen in the posterior for the real data can be at least partially attributed to sample size.

5. Discussion

The first conclusion we draw from this study is that the multiple populations can be inferred from the current sample of hot Jupiters given the limitations of our hierarchical Bayesian framework. This suggests that one could potentially disentangle populations consistent with having a disk or eccentric migration history, perhaps placing informative constraints on the prevalence of each mechanism given a large enough sample size (at least a few hundred). Of course, there are many caveats to address in regards to the data sets, model assumptions, and interpretations of the final results.

5.1. The Data

5.1.1. Choosing the NASA Exoplanet Archive versus exoplanets.org

We present results based on planets listed in the NASA Exoplanet Archive but a similar analysis applied to another exoplanet database provides an important test on how sensitive our results are to the criteria for which planets are included in these data sets.

There are a few differences in the planets queried from the NASA Exoplanet Archive and exoplanets.org when using similar filters. The most substantial difference was in the number of HAT planets: 51 for exoplanets.org and 77 for the Exoplanet Archive. This contains many more HAT South planets than the former, but the cumulative $x$ distributions are very similar. The Exoplanet Archive also finds fewer WASP planets than exoplanets.org. Some WASP planets do not have a listed RV semi-amplitude, and these are missed when we filter for $K > 30$ m s$^{-1}$. Other exoplanets including HD 189733 are also missed by the Exoplanet Archive due to this semi-amplitude filtering issue.

5.1.2. The NASA Exoplanet Archive Data Set

Focusing on the RV planets specifically, we find that $x_{1} \approx 2$ holds even after a decade of new discoveries. The Kepler data also seem to show a similar cutoﬀ, albeit with a smaller sample.

On one hand, this may mean that this subset of Kepler stars and those probed by RV are similar or at least share common conditions of their local galactic environments. On the other hand, the results of our analysis of the HAT+WASP data may suggest that we are missing some very short-period hot Jupiters in the RV or Kepler data. For the latter, it is known that part of the Transit Planet Search software module searches for and removes a set of known harmonics in the flux time series (Jenkins et al. 2010; Tenenbaum et al. 2012). Some very short-period transits may resemble these harmonics and be accidentally removed from these data. If this has been the case, there could be a handful of $1 < x < 2$ hot Jupiters around relatively bright stars awaiting discovery in the Kepler data. Another possibility for the lack of $1 < x < 2$ planets in the RV and Kepler data may stem from insufficient sampling of the true population.

The HAT+WASP data show an increase in planet frequency until $x \approx 3$ then a tapering at higher values. Transit surveys should be mostly biased against detecting planets with larger $x$, so the physical meaning behind the peak around $x \approx 3$ is unclear. If such biases in these long-running surveys are not negligible, we would expect that the mode of this hot-Jupiter population shifts to slightly higher values and the distribution would have a stronger tail toward larger $x$ values.

5.2. The Model

Simple theories of planet formation generally fail to explain the wide diversity of planetary systems seen to date. Many of the model assumptions were motivated more by statistical convenience rather than an accurate reflection the observed hot-Jupiter population(s).

For instance, many of the radius measurements had very asymmetrical errorbars, ranging from small to order of magnitude differences. For simplicity, we modeled the observed radii as being drawn from a symmetric Gaussian distribution. Future work can improve upon this using a more general prescription of modeling asymmetric uncertainties (e.g., using a skewed Gaussian) or using the joint posterior distributions of $P$, $M$, and $R$ if available.

For each data set, we set unknown radii to $1.2 R_{J}$ and $\sigma_{R} = 0.1 R_{J}$. FRO6 found a near perfect degeneracy between $R$ and $x_{1}$ in the RV sample, even when including known transiting planets, so we would expect this relation to hold if we chose a different $R$. In other words, setting unknown radii to a larger (smaller) mean value would simply decrease (increase) $x_{1}$ by a proportional amount.

A major limitation of the truncated Pareto is that $x_{1}$ largely depends on the planet with the smallest $x$. Imposing such a hard edge may not offer enough flexibility with the detailed physics suggested by the observations. Eccentric migrating planets that circularize at $x \approx 2$ may experience further orbital decay due to long-term tidal dissipation within the slowly rotating host star (Valsecchi & Rasio 2014a). This orbital decay timescale can be very sensitive to the planet’s semimajor axis depending on the tidal evolution prescription (e.g., $a^{8}$, Valsecchi & Rasio 2014b). This would smear out the lower truncation and create a low probability tail inward of $x = 2$ that these surveys may not have probed yet. So, for example, if one were to motivate that the observations were consistent with a population with an eroded edge, they could instead model a power law with an inner “soft” edge, perhaps a piecewise function parameterized with a linear slope or exponential decay. This could provide an alternative explanation for the observations and should certainly be investigated in a future study.

Our primary interest here is to look at these data in the context of the two well-studied hot-Jupiter formation channels and lay the groundwork for more complex modeling of single or overlapping populations. Any detailed hot-Jupiter formation theory needs to motivate a particular population-level distribution to sample from, and such an exhaustive investigation of these theories and how to model their representative distributions is beyond the scope of this paper.

5.3. Interpreting the Results

Sampling from the single-component model was computationally tractable and the Markov chains appear to be well
sampled. The two-component model proved to be a much greater computational challenge. At first glance of the parameters in Figure 6, it seems the MCMC suffered from sampling issues in the model that neglected uncertainties, since multiple modes were found in $x_{12}$ space. We find these modes have comparable posterior probabilities, so the data do not seem to favor a unique unimodal solution. Nevertheless, Stan was able to sample from multiple modes quite efficiently, and the Gelman–Rubin statistic shows that multiple chains were able to converge to similar distributions.

Other than the location of the inner cutoffs, we also place constraints on $\gamma_1, \gamma_2$, and $\gamma_3$. For the Kepler and RV samples, we find $\gamma = -0.51 \pm 0.19$ for a single-component Pareto. We compare this to $\gamma_2 = -1.38 \pm 0.32$ for the second component of the two-component mixture model applied to the HAT+WASP samples. These estimates are consistent to within $2\sigma$, but this can be interpreted in a few ways. First, there is an anti-correlation between $x_1$ and $\gamma$, since increasing $x_1$ for a fixed $x_0$ will push $\gamma$ to more negative values. This is most easily seen in the top panel of Figure 5 for the RV data alone, where our estimate of $x_1$ overlaps much better with that of $x_{12}$. In this case, the power-law indices are consistent to within $1\sigma$. Second, perhaps non-negligible observing biases exist in the HAT+WASP samples that negatively affect the detection of planets at larger $x$ and accounting for these would likely increase $\gamma_2$, matching up with the RV+Kepler results much better. In any case, the uncertainty in $\gamma_2$ is moderately large, and it is unclear if the distributions associated with $\gamma$ and $\gamma_2$ probe similar populations of planetary systems.

After applying our statistical models to the exoplanets.org data, we arrive at similar conclusions regarding our population inference. The RV+Kepler data do not find significant evidence for $M_2$ (Bayes factor $\approx 0.1$). For $M_1$, we find $x_1 \approx 2$ and $\gamma = -0.53 \pm 0.23$, which is consistent with the Exoplanet Archive planets. The HAT+WASP data find decisive evidence for $M_2$ (Bayes factor $\approx 10^{14}$) with moderately different $\gamma$ values and mixture fractions. We find that $35 \pm 10\%$ reside in the disk migration component ($\gamma = 0.82 \pm 0.23$) and $65 \pm 10\%$ in the high-eccentricity migration component ($\gamma = -2.45 \pm 0.34$).

Given the evidence for two populations, it is natural to ask whether the mixture fractions correlate with other observable properties, such as the host star temperature, metallicity, or rotation rate. Based on an exploratory analysis of the Exoplanet Archive data set, we identify a potential difference in the $x$ distribution of planets with host star effective temperature. Planets with $1 < x < 2$ seem to be preferentially around hotter stars. We caution that there is a relatively small number of planets in this range and that this apparent difference is not replicated when we perform a similar analysis of the exoplanets.org data set. This demonstrates that, despite enormous progress in these exoplanet surveys, there is still motivation for further increasing the sample size in order to address basic questions about the formation of hot Jupiters.

5.4. Looking Forward

Upcoming exoplanet missions will significantly expand the hot-Jupiter sample.

The KELT survey (Pepper et al. 2007, 2012) has published over a dozen hot Jupiters, with at least one near tidal disruption (Oberst et al. 2017), and is increasing sensitivity to longer-period planets. The Transiting Exoplanet Survey Satellite (TESS, Ricker et al. 2015) is an all-sky mission designed to target nearby bright stars that are amenable to ground-based transit and RV follow-up. TESS will stare at 13 “sectors” over the course of a year, which should yield on the order of 70 hot Jupiters around pre-selected stars observed with a 2-minute time sampling and on the order of 10,000 around stars observed in full-frame images with a 30-minute sampling (Sullivan et al. 2015). However, this survey design means that stars closer to the ecliptic poles will receive a longer observational baseline, enabling longer-period transits to be characterized. This observational bias may need to be accounted for when studying hot Jupiters from TESS, depending on the assumed orbital period cutoff that defines this population. Furthermore, getting mass measurements of every TESS hot Jupiter would be an intractable observational campaign. In principle, one could invoke a probabilistic mass–radius relation (e.g., Chen & Kipping 2017; Wolfgang et al. 2016). Since $R$ depends very weakly on $M$ in the known giant planet regime, this could be offset with a large enough sample size and $x$’s weak dependence on $M$.

Looking into the farther future, WFIRST will measure precise light curves for millions of stars with the intent of finding long-period or free-floating microlensing planets. However, this survey also has an expected yield of several thousand transiting hot Jupiters (Montet et al. 2017). These stars would most likely be too faint to make precise planetary mass measurements. However, the population statistics of these hot Jupiters residing near the Galactic center can be directly compared to studies of those in the field, which can reveal details on hot-Jupiter formation as a function of local galactic environment.

All of the results and code presented in this project are available at https://github.com/benelson/hjs_with_stan and archived at Zenodo (doi:10.5281/zenodo.834744). This repository includes results for both the NASA Exoplanet Archive and exoplanets.org data sets.

We thank Joseph Catanzarite, Adam Dempsey, Sybil Derribe, Sam Hadden, Jason Hwang, Yoram Lithwick, Daniel Stevens, Francesca Valsecchi, and Angie Wolfgang for helpful discussions and contributions. This research has made use of the NASA Exoplanet Archive, which is operated by the California Institute of Technology, under contract with the National Aeronautics and Space Administration under the Exoplanet Exploration Program. This research has also made use of the Exoplanet Orbit Database and the Exoplanet Data Explorer at exoplanets.org.

B.E.N. acknowledges support from the Data Science Initiative at Northwestern University. E.B.F. acknowledges support from NASA Grant NNX14AN77G and NNX14A177G. F.A.R. acknowledges support from NASA Grant NNX12AI86G at Northwestern University and from NSF Grant PHY-1066293 through the Aspen Center for Physics. This research was supported in part through the computational resources and staff contributions provided for the Quest high performance computing facility at Northwestern University, which is jointly supported by the Office of the Provost, the Office for Research, and Northwestern University Information Technology. This material was based upon work partially supported by the National Science Foundation under Grant DMS-1127914 to the Statistical and Applied Mathematical Sciences Institute. Any opinions, findings, and conclusions or recommendations

10
expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.  

Software: daft (http://daft-pgm.org), NFA (Derrible & Ahmad 2015), Pystan (Stan Development Team 2017), Stan (Carpenter et al. 2017).

ORCID iDs

Benjamin E. Nelson @ https://orcid.org/0000-0003-3010-2334
Eric B. Ford @ https://orcid.org/0000-0001-6545-639X
Frederic A. Rasio @ https://orcid.org/0000-0002-7132-418X

References

Aban, I. B., Meerschaert, M. M., & Panorska, A. 2006, J. Am. Stat. Assoc., 101, 270
Akeson, R. L., Chen, X., Ciardi, D., et al. 2013, PASP, 125, 989
Bakos, G., Noyes, R. W., Kovács, G., et al. 2004, PASP, 116, 266
Beg, M. 1981, Journal of Information and Optimization Sciences, 2, 192
Bryan, M. L., Knutson, H. A., Howard, A. W., et al. 2016, ApJ, 821, 89
Carpenter, B., Gelman, A., Hoffman, M., et al. 2017, Journal of Statistical Software, 76, 1
Chen, J., & Kipping, D. M. 2017, ApJ, 834, 17
Cumming, A. 2004, MNRAS, 354, 1165
Dawson, R. I., & Murray-Clay, R. A. 2013, ApJL, 767, L24
Derrible, S., & Ahmad, N. 2015, PLoSO, 10, 1
Fabrycky, D., & Tremaine, S. 2007, ApJ, 669, 1298
Ford, E. B., & Rasio, F. A. 2006, ApJL, 638, L45
Foreman-Mackey, D., Hogg, D. W., & Morton, T. D. 2014, ApJ, 795, 64
Goldreich, P., & Tremaine, S. 1980, ApJ, 241, 425
Han, E., Wang, S. X., Wright, J. T., et al. 2014, PASP, 126, 827
Hoffmann, M. D., & Gelman, A. 2014, Journal of Machine Learning Research, 15, 1351
Jenkins, J. M., Chandrasekaran, H., McCauliff, S. D., et al. 2010, Proc. SPIE, 7740, 7740D
Knutson, H. A., Fulton, B. J., Montet, B. T., et al. 2014, ApJ, 785, 126
Lieu, M., Farr, W. M., Betancourt, M., et al. 2017, MNRAS, 468, 4972
Lin, D. N. C., Bodenheimer, P., & Richardson, D. C. 1996, Natur, 380, 606
Mizuno, H. 1980, PThPh, 64, 544
Montet, B. T., Yee, J. C., & Penny, M. T. 2017, PASP, 129, 044401

Murray, N., Hansen, B., Holman, M., & Tremaine, S. 1998, Sci, 279, 69
Naoz, S., Farr, W. M., Lithwick, Y., Rasio, F. A., & Teyssandier, J. 2011, Natur, 473, 187
Naoz, S., Farr, W. M., & Rasio, F. A. 2012, ApJL, 754, L36
Nelson, B. E., Robertson, P. M., Payne, M. J., et al. 2016, MNRAS, 455, 2484
Neveu-VanMalle, M., Queloz, D., Anderson, D. R., et al. 2014, A&A, 572, A49
Ngo, H., Knutson, H. A., Hinkley, S., et al. 2015, ApJL, 800, 138
Ngo, H., Knutson, H. A., Hinkley, S., et al. 2016, ApJL, 827, 8
Oberst, T. E., Rodriguez, J. E., Colón, K. D., et al. 2017, AI, 153, 97
Pepper, J., Kuhn, R. B., Siverd, R., James, D., & Stassun, K. 2012, PASP, 124, 230
Pepper, J., Poppe, R. W., DePoy, D. L., et al. 2007, PASP, 119, 923
Piskorz, D., Knutson, H. A., Ngo, H., et al. 2015, ApJL, 814, 148
Pollack, J. B., Hubickyj, O., Bodenheimer, P., et al. 1996, Ice, 124, 62
Rasio, F. A., & Ford, E. B. 1996, Sci, 274, 954
Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, JATIS, 1, 014003
Rogers, L. A. 2015, ApJ, 801, 41
Rogers, T. M., Lin, D. N. C., & Lau, H. H. B. 2012, ApJL, 758, L6
Sanders, N. E., Betancourt, M., & Soderberg, A. M. 2015, ApJ, 800, 36
Santerne, A., Moutou, C., Tsantaki, M., et al. 2016, A&A, 587, A64
Schlaufman, K. C., & Winn, J. N. 2016, ApJL, 825, 62
Shabram, M., Demory, B.-O., Cisewski, J., Ford, E. B., & Rogers, L. 2016, ApJ, 820, 93
Stan Development Team 2017, Pystan: the Python Interface to Stan, http://mc-stan.org, v2.16.0.0
Street, R. A., Pollacco, D. L., Fitzsimmons, A., et al. 2003, in ASP Conf. Ser. 294, Scientific Frontiers in Research on Extrasolar Planets, ed. D. Deming & S. Seager (San Francisco, CA: ASP), 405
Sullivan, P. W., Winn, J. N., Berta-Thompson, Z. K., et al. 2015, ApJ, 809, 77
Tenenbaum, P., Christiansen, J. L., Jenkins, J. M., et al. 2012, ApJS, 199, 24
Valsecchi, F., Rappaport, S., Rasio, F. A., Marchant, P., & Rogers, L. A. 2015, ApJ, 813, 101
Valsecchi, F., & Rasio, F. A. 2014a, ApJL, 787, L9
Valsecchi, F., & Rasio, F. A. 2014b, ApJ, 786, 102
Valsecchi, F., Rasio, F. A., & Steffen, J. H. 2014, ApJL, 793, L3
Wang, J., Fischer, D. A., Horch, E. P., & Huang, X. 2015, ApJ, 799, 229
Ward, W. R. 1997, Icar, 126, 261
Wolfgang, A., & Lopez, E. 2015, ApJ, 806, 183
Wolfgang, A., Rogers, L. A., & Ford, E. B. 2016, ApJ, 825, 19
Wu, Y., & Lithwick, Y. 2011, ApJ, 735, 109
Wu, Y., & Murray, N. 2003, ApJ, 589, 605
Zakamska, N. L., Pan, M., & Ford, E. B. 2011, MNRAS, 410, 1895