Draft version November 14, 2017
Typeset using LaTeX **modern** style in AASTeX61

# THE EFFECTS OF BINARITY ON PLANET OCCURRENCE RATES MEASURED BY TRANSIT SURVEYS

L. G. Bouma,[1] J. N. Winn,[1] and K. Masuda[1]

[1] *Department of Astrophysical Sciences, Princeton University, 4 Ivy Lane, Princeton, NJ 08540, USA*

Submitted to AAS journals.

## ABSTRACT

This work aims to clarify the biases that stellar binarity introduces to occurrence rates inferred from transit surveys. In general, stellar multiplicity leads to diluted planetary radii, overestimated detection efficiencies, and an undercounted number of selected stars (but possibly an overcounted number of searched stars). These effects skew occurrence rate measurements in different directions, and we develop simple models that allow us to understand the crucial effects. For a model in which all stellar systems are either single or twin binaries, and all planets are identical, we find that ignoring binarity leads to an underestimate of the occurrence at the true radius by a multiplicative factor of 0.76 (assuming a 10% twin binary fraction). Using more realistic models for the stellar population and planetary radii, we find that ignoring binarity leads to a 10 - 30% underestimate of the number of planets per star, depending on the true radius distribution. For the occurrence of Earth-sized planets, in our most realistic model the rate is underestimated by $\approx 15\%$ – at present far smaller than systematic uncertainties on real $\eta_\oplus$ measurements. For hot Jupiters, we find that the inferred occurrence rate is $\approx 1.3\times$ smaller than the true rate around single stars. We suggest that this latter effect contributes to the discrepant hot Jupiter rates measured by *Kepler* and the California group's respective surveys.

*Keywords:* methods: data analysis — planets and satellites: detection — surveys

Corresponding author: L. Bouma
luke@astro.princeton.edu

## 1. INTRODUCTION

A group of astronomers wants to measure the mean number of planets of a certain type per star of a certain type. Ignoring stellar multiplicity, they perform a signal-to-noise limited transit survey and detect $N_{\rm det}$ signals that seem to be from the desired class of planet. They count the apparent number of selected stars, $N_\star$. Assuming their survey is limited by Poisson noise, the group computes their detection efficiency $p_{\rm det}$ as a function of planetary and stellar properties, in order to know which of the selected stars were searchable. Correcting for the geometric transit probability $p_{\rm tra}$, they report an apparent occurrence rate $\Lambda_{\rm a}$:

$$\Lambda_{\rm a} = \frac{N_{\rm det}}{N_\star} \times \frac{1}{p_{\rm tra}} \times \frac{1}{p_{\rm det}}. \tag{1}$$

There are many potential pitfalls. Some genuine transit signals can be missed by the detection pipeline. Some apparent transit signals are spurious, from noise fluctuations, failures of 'detrending', or instrumental effects. Stars and planets can be misclassified due to statistical and systematic errors in the measurements of their properties. Poor angular resolution causes false positives due to blends with background eclipsing binaries. *Et cetera.*

Here we focus on problems that arise from the fact that many stars exist in multiple star systems. For simplicity, we consider only binaries, and we assume that they are all spatially unresolved.

An immediate complication is that, due to dynamical stability or some aspect of planet formation, the occurrence rate of planets might differ between binary and single-star systems. If "occurrence rate" is defined as the mean number of planets within set radius and period bounds per star in a given mass interval, it must implicitly marginalize over stellar multiplicity. This means marginalizing over "occurrence rates in single star systems", "occurrence rates about primaries", and "occurrence rates about secondaries" (see *e.g.,* Wang et al., 2015).

Outside of astrophysical differences, there are observational biases for every term in Eq. 1. There are errors in $N_{\rm det}$ due to planet radius misclassification. There are errors in $N_\star$ because a given selected star might in fact be two stars. There are errors in $p_{\rm tra}$ because stars in binaries may have different masses and radii than assumed in the single-star case. Finally, there are errors in $p_{\rm det}$ because detection efficiencies for planets orbiting single stars, primaries, and secondaries are all different.

Correcting for binarity's observational biases is non-trivial. For instance, in counting the number of selected stars, even after realizing that binaries count as two stars, one must note that the multiplicity fraction of *selected* stars is greater than that of a volume limited sample. This is the familiar Malmquist bias: binaries are selected out to larger distances than single stars because they are more luminous. As a separate challenge, finding the correct number of detected planets in a radius bin, $N_{\rm det}$, requires knowledge of the true planetary radii. Observers deduce apparent radii. In binaries,

the apparent and true radii differ because of diluting flux, and possibly because the planet is assumed to orbit the wrong star (*e.g.*, Furlan et al. 2017).

To gain intuition for the many observational biases at play, we consider a set of idealized transit surveys:

- Model #1: fixed stars, fixed planets, twin binaries;
- Model #2: fixed planets and primaries, varying secondaries;
- Model #3: fixed primaries, varying planets and secondaries.

We define our terminology in Sec. 2, and present our transit survey models in Secs. 3.1-3.3, where each subsection corresponds to a model listed above. We interpret these calculations throughout, and in Sec. 4 connect them to topical questions in the interpretation of transit survey occurrence rates. In particular, we mention the "hot Jupiter rate discrepancy", the relevance towards measurements of $\eta_\oplus$, and the significance for the dearth of planets recently discovered by Fulton et al. (2017). We conclude in Sec. 5

## 2. DEFINITIONS

We define the occurrence rate density, $\Gamma$, as the expected number of planets per star per bin of planetary or stellar phase space. Since we will mainly be concerned with the rate density's dependence on planetary radii $r$, we write

$$\Gamma(r) = \frac{\mathrm{d}\Lambda}{\mathrm{d}r}, \tag{2}$$

where $\Lambda$ is the occurrence rate. In this notation, "the occurrence rate of planets of a particular size" translates to an integral of Eq. 2, evaluated over a radius interval.

The above definition implicitly marginalizes the rate density over stellar multiplicity. In this study we only consider single and binary star systems. For a selected population of stars and planets, the rate density is then a weighted sum of rate densities for each system type:

$$N_{\mathrm{tot}}\Gamma(r) = N_0\Gamma_0(r) + N_1\Gamma_1(r) + N_2\Gamma_2(r), \tag{3}$$

where $i = 0$ corresponds to single stars, $i = 1$ to primaries of binaries, and $i = 2$ to secondaries of binaries. $N_{\mathrm{tot}} = \sum_i N_i$ is the total number of selected stars, and $N_0, N_1, N_2$ are the number of selected single stars, primaries, and secondaries. Since each selected binary system contributes both a primary and secondary star, $N_1 = N_2$. This redundancy in our notation will later prove its use.

Finally, it is helpful to write $\Gamma_i(r)$, the rate density for each type of star, as the product of a shape function and a constant:

$$\Gamma_i(r) = \frac{\mathrm{d}\Lambda_i}{\mathrm{d}r} = Z_i p_i(r), \quad \text{for } i \in \{0, 1, 2\}. \tag{4}$$

The shape function is normalized to unity. The $Z_i$'s can be interpreted as each system type's occurrence rate $\Lambda_i$, integrated over all planetary radii. They are equivalent to the number of planets per single star, primary, or secondary.

## 3. IDEALIZED MODELS OF TRANSIT SURVEYS

### 3.1. *Model #1: fixed stars, fixed planets, twin binaries*

Since the effects of binarity are most pronounced when the two components are similar, we begin by considering a universe in which all planets are identical, and all stars are identical except that some fraction of them exist in binaries.

Expressed mathematically, from Eqs. 3 and 4 the occurrence rate density at a planet radius $r$ can be written

$$\Gamma(r) = \delta(r_p) \times \frac{N_0 Z_0 + N_1 Z_1 + N_2 Z_2}{N_{\rm tot}}, \tag{5}$$

where $\delta(r_p)$ is the Dirac delta function, zero except at the true planet radius, $r_p$. The occurrence rate over any interval that includes $r_p$ is then

$$\Lambda|_{r_p} = \frac{N_0 Z_0 + N_1 Z_1 + N_2 Z_2}{N_{\rm tot}}, \tag{6}$$

and the rate is zero over intervals that do not include $r_p$.

We return to our group of binarity-ignoring astronomers. They do not know the true rate density – they would like to discover it! In their signal-to-noise limited transit survey, they select stars that they think can yield transit detections. Since the noise is Poissonian, they assume

$$\frac{\rm signal}{\rm noise} \propto \frac{(r/R)^2}{F^{-1/2}} \propto F^{1/2} \propto L_{\rm sys}^{1/2} d^{-1}, \quad \text{(incorrectly assumed)} \tag{7}$$

for $R$ the constant stellar radius, $F$ the photon flux, $L_{\rm sys}$ the luminosity of a system, and $d$ its distance from us. At fixed planet radius, semimajor axis, stellar radius, and stellar luminosity, a constant signal-to-noise floor yields a maximum detectable distance (Pepper et al 2003, Pepper and Gaudi 2005). The maximum distance out to which our binarity-ignoring astronomers select stars, $d_{\rm sel}$, thus scales as $L_{\rm sys}^{1/2}$.

The single stars have luminosity $L_1$, and the twin binaries have luminosity $2L_1$. Thus the twin binaries are selected out to a distance $\sqrt{2}$ times that of single stars. This is a bad move, because the transit signal for any planet in a twin binary will be diluted by a factor of two:

$$\frac{\rm signal}{\rm noise} \propto \frac{\mathcal{D}(r/R)^2}{F^{-1/2}} \propto \mathcal{D} F^{1/2} \propto \mathcal{D} L_{\rm sys}^{1/2} d^{-1}, \quad \text{(true)} \tag{8}$$

where the dilution is $\mathcal{D} \equiv L_{\rm host}/L_{\rm sys}$, for $L_{\rm host}$ the the planet host's luminosity. This means that the true maximum searchable distance for binaries, $d_{\rm det}$, is $1/\sqrt{2}$ times that of single stars. The situation is illustrated in Fig. 1: only one in eight selected stars in binaries are truly searchable.

*What do the observers ignoring binarity infer?*—The binarity-ignoring observers assume that all points on the sky with flux above some minimum are searchable. They correct

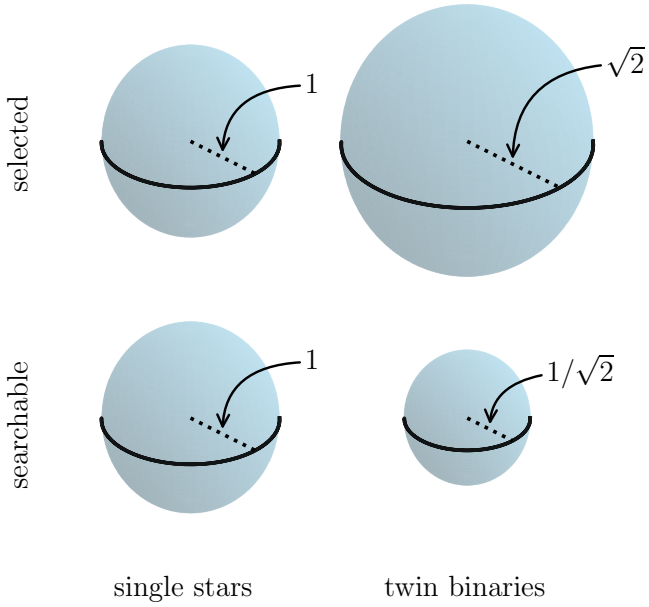


**Figure 1.** Cartoon of the different selected and searchable volumes for single stars (left column), and twin binaries (right column). This model (#1) assumes all stars have equal mass and luminosity. Since the observer does not recognize the brighter binaries, they select them out to a larger distance, $\sqrt{2}\times$ that of single stars. However, dilution causes the twin binaries to only be searchable in one eighth of their selected volume.

their assumed number of searchable stars for the transit probability. As a function of apparent radius, they then report an apparent rate density of

$$\Gamma_a(r_a) = \delta(r_p) Z_0 \frac{N_0}{N_0 + N_1} + \delta\left(\frac{r_p}{\sqrt{2}}\right)(Z_1 p_{\rm det,1} + Z_2 p_{\rm det,2}) \frac{N_1}{N_0 + N_1}, \tag{9}$$

where $p_{\rm det,1}$ ($p_{\rm det,2}$) is the probability that a selected primary (secondary) is searchable. In this example, $p_{\rm det,1} = p_{\rm det,2} = 1/8$; the detection efficiency is the ratio of the searchable to selected volumes.

The true rate density (Eq. 5) and the apparent rate density (Eq. 9) differ in that

1. The total number of selected stars, $N_{\rm tot} = N_0 + N_1 + N_2$, was miscounted.
2. The detection efficiency was incorrectly assumed to be 1 for all selected stars. In reality, only one in eight binaries were searchable.
3. The inferred radii in binary systems are all $\sqrt{2}$ too small.

To assess the severity of these errors, we need to make assumptions about the stellar and planetary populations. The important quantity for the stellar population is $N_1/N_0$, the ratio of selected primaries (or binaries) to singles. We can define

$$\mu \equiv \frac{N_1}{N_0} = \frac{n_b}{n_s}\left(\frac{d_{\rm sel,b}}{d_{\rm sel,s}}\right)^3 = \frac{\rm BF}{1 - \rm BF}(1 + \ell)^{3/2}, \tag{10}$$

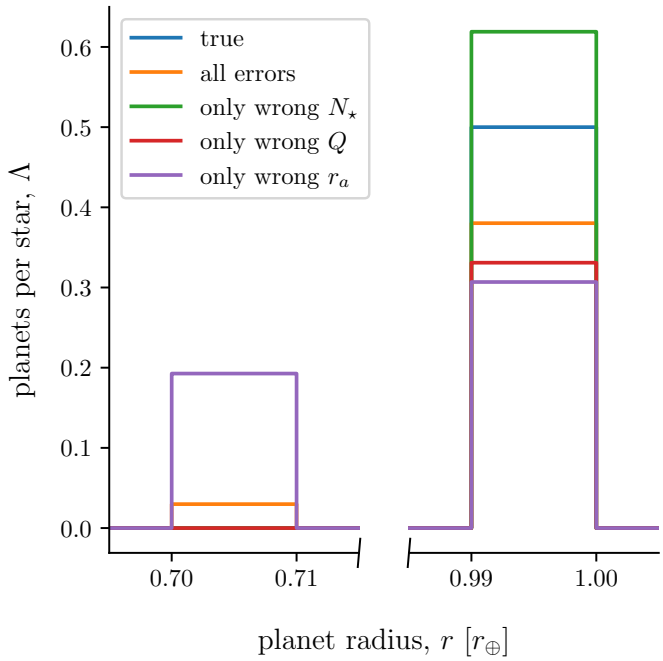


**Figure 2.** Inferred planet occurrence rates over $0.01 r_\oplus$ bins in planet radius, for Model #1. This model has fixed stars, fixed planets, and twin binaries. It assumes a twin binary fraction of BF = 0.1. If the true planet radius is $r_p$, all planets detected in binaries will have apparent radii $r_a = r_p/\sqrt{2}$. We illustrate the individual biases by separating them. On this and future plots, "only wrong $N_\star$" means the only error is an incorrectly assumed number of selected stars; "only wrong $Q$" means the only error is an incorrectly assumed completeness (including both miscalculated $p_{\rm tra}$ and fraction of selected stars that are searchable, $p_{\rm det}$); "only wrong $r_a$" means the only error is in miscalculated planetary radii, due to both transit depth dilution and also wrongly assumed host star radii.

where $n_b$ and $n_s$ are the number density of binaries and singles in a volume limited sample, and the maximum selected distance for binary and single systems are $d_{\rm sel,b}$ and $d_{\rm sel,s}$. The light ratio $\ell$ is defined relative to the primary, so that $L_{\rm sys} = L_1(1+\ell)$. The binary fraction in a volume limited sample is "BF". For instance, Raghavan et al., (2010) found a multiplicity fraction[1] of 0.44 for primaries with masses from $0.7 M_\odot$ to $1.3 M_\odot$. "Twin binaries" are perhaps 10-20% of the binaries in a volume-limited sample, depending on how "twin" is defined. Thus a representative twin binary fraction for this model is BF $\approx$ 0.05-0.1.

Further, we need to assume knowledge of the true number of planets per single, primary, and secondary star ($Z_0, Z_1$, and $Z_2$). Taking $Z_0 = Z_1 = Z_2 = 1/2$, we plot the resulting occurrence rates as a function of planet radius in Fig. 2. Though it is impossible for an observer to isolate any one of binarity's biases without also addressing the others, mathematically it is simply a matter of modifying the appropriate

[1] The binary fraction is the fraction of systems in a volume-limited sample that are binary. It is equivalent to the multiplicity fraction if there are no triple, quadruple, or higher order multiples. In that case, BF $= n_b/(n_s + n_b)$.

terms in Eq. 9. The resulting rates are also shown in Fig. 2, and demonstrate how each error individually affects the overall inferred rate.

*Correction to inferred rate density and inferred rate*—We define a rate density correction factor, $X_\Gamma$, as the ratio of the apparent to true rate densities:

$$X_\Gamma \equiv \frac{\Gamma_a}{\Gamma}. \tag{11}$$

This correction factor can be a function of whatever parameters $\Gamma_a$ and $\Gamma$ depend on; in this study, the planet radius is most relevant.

If we continue assuming that the number of planets per single, primary, and secondary star are equal ($Z_0 = Z_1 = Z_2$), we find a rate density correction factor at the true planet radius of

$$X_\Gamma(r_p) = \frac{1}{1+\mu}, \tag{12}$$

This yields a correction of 0.76 if BF = 0.1, and 0.87 if BF = 0.05. Rephrasing the result, if the twin binary fraction were $(2^{3/2}+1)^{-1} \approx 0.26$, then the apparent rate would be half the true rate. Fortunately, in most contexts the twin binary fraction is not that high.

An alternative assumption is that secondaries do not host planets. In that case, $Z_0 = Z_1$, and $Z_2 = 0$. The correction to the rate density at the true planet radius becomes

$$X_\Gamma(r_p) = \frac{1+2\mu}{(1+\mu)^2}. \tag{13}$$

This evaluates to 0.94 if BF = 0.1, and 0.98 if BF = 0.05. While it is hard to justify the assumption that secondaries are planet-less, it is worth noting that $\Gamma_a/\Gamma$ is sensitive to the relative number of planets per single, primary, and secondary.

### 3.2. *Model #2: fixed planets and primaries, varying secondaries*

Our binary-twin model provides a simple estimate of binarity's systematic effects, but perhaps it is too simple. We begin introducing realism by first letting the light ratio $\ell = L_2/L_1$ vary across the binary population. It does so because the underlying mass ratio $q = M_2/M_1$ varies. We keep the primary mass fixed as $M_1$, which is also the mass of all single stars.

We parametrize the distribution of binary mass ratios in a volume-limited sample as a power law: $p(q) \propto q^\beta$. For binaries with solar-type primaries[2], $\beta$ is probably between 0 and 0.3. We further assume that stars are a one-parameter family, $R \propto M \propto L^{1/\alpha}$, so that a drawn value of $q$ determines everything about a secondary.

The rate density in this model, $\Gamma(r, q)$, is the sum of the rate densities for each system type:

$$\Gamma(r, q) = \delta(r_p) \times \frac{N_0 Z_0 + N_1 Z_1 + N_2 Z_2 p_2(q)}{N_{\rm tot}} \tag{14}$$

[2] Duchene and Kraus (2013), fitting all the multiple systems of Raghavan et al. (2010)'s Fig 16, find $\beta = 0.28 \pm 0.05$ for $0.7 < M_\star/M_\odot < 1.3$. Examining only the binary systems of Rhagavan et al 2010, Fig 16, the distribution seems roughly uniform, $\beta \approx 0$, except for a claimed excess of twin binaries with $q \approx 1$, and an obvious lack of $q < 0.1$ stellar companions.

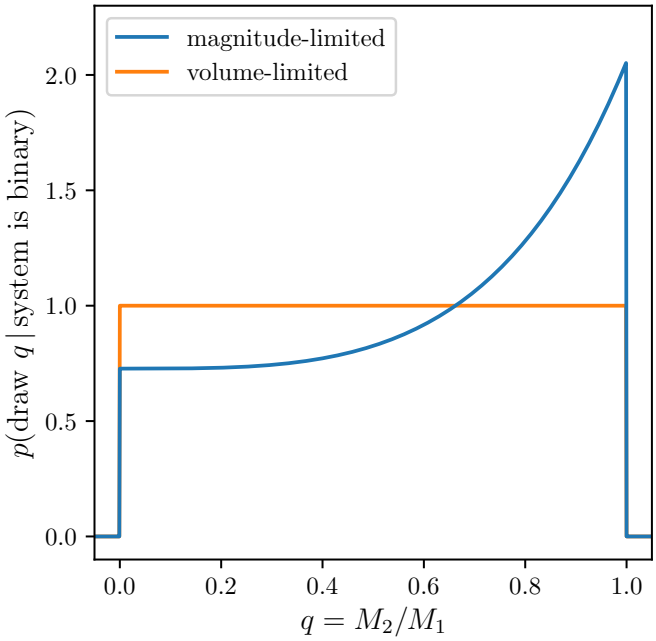


**Figure 3.** The mass ratio distribution for a magnitude-limited sample of binary stars, in which the underlying volume-limited distribution is uniform (quite similar to *e.g.*, Raghavan et al. (2010)'s Fig. 16). The entire bias can be understood analytically (Eq. 15).

where $p_2(q)$ is mass-ratio dependent shape function in secondaries. A fully general approach would parametrize $p_2(q)$ as a power law, $p_2(q) \propto q^\gamma$. For simplicity we will always take $\gamma = 0$, the uniform case. The interpretation of $Z_i$ is still "the number of planets per star of type $i$", but now for secondaries this must include a marginalization over both the mass ratio and the planet radius.

There is also a Malmquist bias in Eq. 14, hidden away in the numbers of selected stars. This is because the selected sample at a fixed planet radius and period is magnitude-limited. Given a binary, the probability of drawing a mass ratio $q$ in a magnitude-limited sample scales as

$$p(\text{draw } q \,|\, \text{system is binary}) \propto q^\beta (1 + q^\alpha)^{3/2} \tag{15}$$

where $q^\beta$ is the volume-limited probability of drawing a binary of mass ratio $q$, and the latter term is the Malmquist bias. We show the magnitude-limited mass ratio distribution for the $\beta = 0$ case in Fig. 3. We emphasize that in Monte Carlo simulations of transit surveys, it is important to draw binaries from the correctly biased mass-ratio distribution (e.g., Bakos et al 2012, Sullivan et al 2015, Guenther et al 2017).

The occurrence rate corresponding to Eq. 14's rate density for specific mass ratios of interest $q_{\rm min} < q < q_{\rm max}$ is then

$$\Lambda|_{r_p, q_{\rm min}, q_{\rm max}} = \frac{N_0 Z_0 + N_1 Z_1 + N_2 Z_2 f_2}{N_{\rm tot}}, \tag{16}$$

for

$$f_2 \equiv \int_{q_{\min}}^{q_{\max}} p_2(q)\,\mathrm{d}q, \tag{17}$$

where $p_2(q) \propto q^\gamma$, and must be normalized to unity.

*What do the observers ignoring binarity infer?*—As a reminder, the binary-ignoring observers compute an apparent rate density by correcting the number of detections per star for the transit probability and the detection efficiency. In this process, they make the following errors:

1. They assume that they have selected $N_0 + N_1$ stars, while they have actually selected $N_0 + N_1 + N_2$.

2. They assume that every star within the maximum selected distance is searchable. This is true for single stars. For binaries, the actual detection efficiency is the ratio of the number of searchable stars to the number of selected stars. For primaries,

$$p_{\mathrm{det},1} = \left(\frac{d_{\mathrm{det},1}}{d_{\mathrm{sel}}}\right)^3 = \mathcal{D}_1^3 = (1+q^\alpha)^{-3}, \tag{18}$$

   where we have assumed that all single stars and primaries have the same properties, and have used $\ell = q^\alpha$. For secondaries,

$$p_{\mathrm{det},2} = \left(\frac{d_{\mathrm{det},2}}{d_{\mathrm{sel}}}\right)^3 = \mathcal{D}_2^3 \left(\frac{R_1}{R_2}\right)^6 \left(\frac{T_{\mathrm{dur},2}}{T_{\mathrm{dur},1}}\right)^{3/2} = (1+q^{-\alpha})^{-3} q^{-5}, \tag{19}$$

   where subscripts '1' and '2' correspond to primaries and secondaries, and we used the scaling relation $T_{\mathrm{dur}} \propto \rho_\star^{-1/3} \propto R_\star^{2/3}$.

3. They assume a transit probability for all stars of $p_{\mathrm{tra}} = R_\star/a$. For planets orbiting secondaries at fixed period, the true transit probability is $q^{2/3} p_{\mathrm{tra}}$, since the secondaries have smaller radii and masses.

4. The true planetary radii $r$ are interpreted as apparent radii $r_a$. In binaries, the apparent radii depend on whether the host is the primary or secondary:

$$r_a = \begin{cases} r_p(1+q^\alpha)^{-1/2} & \text{for } i=1, \text{ primary} \\ r_p(1+q^{-\alpha})^{-1/2} q^{-1}, & \text{for } i=2, \text{ secondary.} \end{cases} \tag{20}$$

   The factor of $q^{-1}$ for the secondary case accounts for the observer assuming all transit signals come from stars of fixed size.

To write the apparent rate density as a function of the apparent radius $r_a$, we marginalize out the planet period, semimajor axis, and stellar radius (or equivalently the mass ratio, for binaries):

$$\Gamma_a(r_a) = \frac{N_0}{N_0+N_1} Z_0 \delta(r_p) + \frac{N_1}{N_0+N_1} \left(Z_1 I_1(r_a) + Z_2 I_2(r_a)\right). \tag{21}$$

The ratio of primaries to singles, $\mu = N_1/N_0$, is now less than that of Model #1 (cf. Eq. 10). This is because a distribution of light ratios $\ell$ leads to a distribution of maximum selected distances. Integrating over the mass ratios from $q = 0$ to 1, one finds a dimensionless integral

$$\mu \equiv \frac{N_1}{N_0} = \frac{\mathrm{BF}}{1-\mathrm{BF}} \left( 2^{3/2} - \int_1^{\sqrt{2}} u^2 (u^2-1)^{1/\alpha} \, \mathrm{d}u \right). \tag{22}$$

Given a binary fraction, Eq. 22 fully specifies the "weights" in Eq. 21. The $I_1(r_a)$ and $I_2(r_a)$ terms are found by marginalizing over the joint distribution of apparent radius and mass ratio:

$$I_i(r_a) = \int_0^1 p(\text{has detected planet}, r_a, q | \text{star is type } i) \, \mathrm{d}q, \quad \text{for } i \in \{1,2\}, \tag{23}$$

$$= \int_0^1 p(\text{has detected planet} | r_a, q, \text{star is type } i) \times p(r_a | q, \text{star is type } i) \, p(q | \text{star is type } i) \, \mathrm{d}q. \tag{24}$$

The first term is the detection efficiency; the second is a $\delta$-function of the apparent radius; the last is the mass ratio distribution given by Eq. 15. An analytic solution can be found for $i = 1$. For $i = 2$ there is no analytic solution, because evaluating the integral requires imposing the constraint that $r_a = r_p(1+q^{-\alpha})^{-1/2} q^{-1}$. This equation can be re-written

$$\left( \frac{r_p}{r_a} \right)^2 = q^2 + q^{-\alpha+2}, \tag{25}$$

which has no analytic solution for $q(r_a)$ except for special values of $\alpha$, the mass-luminosity exponent. For $\alpha = 3.5$, our nominal case, semianalytic solutions do exist. However, since our main interest is in understanding the qualitative behavior of the solutions, we derive a limiting case, and then proceed numerically.

*Limiting case of rate density correction*—Recall that the rate density correction factor, $X_\Gamma$, is the ratio of the apparent to true rate densities. We consider a "nominal model" in which the stellar population is similar to Sun-like stars in the local neighborhood: BF = 0.44, $\alpha = 3.5$, $\beta = 0$. Our default assumption is also that the occurrence of planets is independent of stellar mass ($\gamma = 0$), so secondaries have the same occurrence rate as primaries and single stars. Under these assumptions, the true rate density is

$$\Gamma(r) \approx \delta(r_p) \left( Z_0 + Z_1 + Z_2 \right) / 3, \tag{26}$$

where the coefficients of 1/3 are accurate to within one percent of the true coefficients. Ignoring binarity, the observer finds an apparent rate density

$$\Gamma_a(r) = c_0 Z_0 \delta(r_p) + c_1 Z_1 I_1(r_a) + c_2 Z_2 I_2(r_a), \tag{27}$$

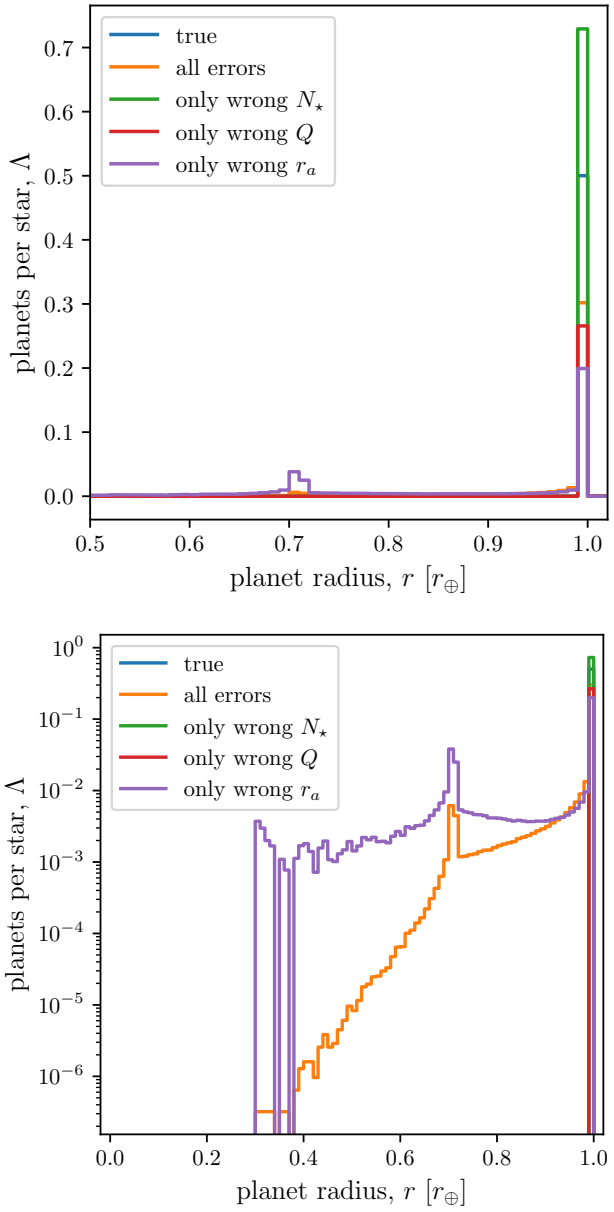


**Figure 4.** Inferred planet occurrence rates as a function of planet radius in Model #2 (*top:* linear, *bottom:* logarithmic). This model has fixed planets and primaries, and varying secondary masses, radii, and luminosities. The latter three lines are described in Fig. 2.

for $c_0 \approx 0.49$, and the coefficients $c_1, c_2$ unknown. We can evaluate the correction term at $r = r_p$, since $\lim_{r_a \to r_p} I_i(r_a) = 0$ for $i \in \{1, 2\}$:

$$X_\Gamma(r = r_p) \approx \frac{3c_0\Lambda_0}{\Lambda_0 + \Lambda_1 + \Lambda_2}. \tag{28}$$

If all the $Z_i$'s are equal, $X_\Gamma(r = r_p) \approx 0.49$. If there are no planets around the secondaries, $X_\Gamma(r = r_p) \approx 0.74$.

*Numerical approach*—To maintain simplicity, we develop a Monte Carlo program to simulate our toy transit surveys. The program generates a stellar population,

assigns planets to the stars, and then calculates which planets are detectable. It then computes the apparent and true planet occurrence rates as a function of planet radius. This is discussed in detail below, and our implementation is available online[3].

First, the user must specify the free parameters that describe the stellar and planetary population. These values include the binary fraction, and the true planet occurrence rates around single stars, primaries, and secondaries (the $Z_i$'s; recall Eq. 4). There is also an arbitrary absolute number of selected stars.

Once the user specifies the free parameters, the program constructs a population of selected stars. Each selected star is assigned a type (single, primary, secondary), a binary mass ratio (if it is not single), and the property of whether it is "searchable". The relative number of binaries to primaries is set by Eq. 22. The mass ratios are drawn from the appropriate magnitude-limited distribution (Eq. 15). If it is single, a selected star is assumed to be searchable. If it is a primary or secondary, its searchability is determined by a uniform draw from either Eq. 18 or Eq. 19, as appropriate.

Assigning planets, each selected star receives a planet at the rate $Z_i$, according to its type. The radii of planets are assigned independently of any host system property, and can be sampled for an arbitrary distribution (*e.g.*, power law; delta function). The absolute transit probability, $p_{\rm tra}$ is set arbitrarily, and the probability of a planet transiting a secondary is set as $q^{2/3} \times p_{\rm tra}$. A planet is "detected" when a) it transits, and b) its host star is searchable. For detected planets, apparent radii are computed according to analytic formulae that account for both dilution and the misclassification of stellar radii (Eq. 20). We assume that the observers think that all transits are around single stars.

We then compute rates in bins of true planet radius and apparent planet radius. In a given radius bin, the true rate is found by counting the number of planets that exist around selected stars of all types (singles, primaries, secondaries), and dividing by the total number of these stars. The apparent rates are found by counting the number of detected planets that were found in an apparent radius bin, dividing by the geometric transit probability for single stars, and dividing by the apparent total number of stars.

*Numerical results for Model #2*—We first validate our model by ensuring it produces the analytically-predicted results for Model #1, and the limiting case of Model #2 described above. Following validation, assuming BF $= 0.44$, $\alpha = 3.5$, $\beta = \gamma = 0$, and taking $Z_0 = Z_1 = Z_2 = 0.5$, we compute the true and apparent occurrence rates over bins in planet radius. The results are shown in Fig. 4. Evidently, dilution produces a spectrum of apparent planetary radii. This leads to overestimated rates everywhere except where there are actually planets, where the rate is underestimated by a factor of two.

[3] https://github.com/lgbouma/binary_biases

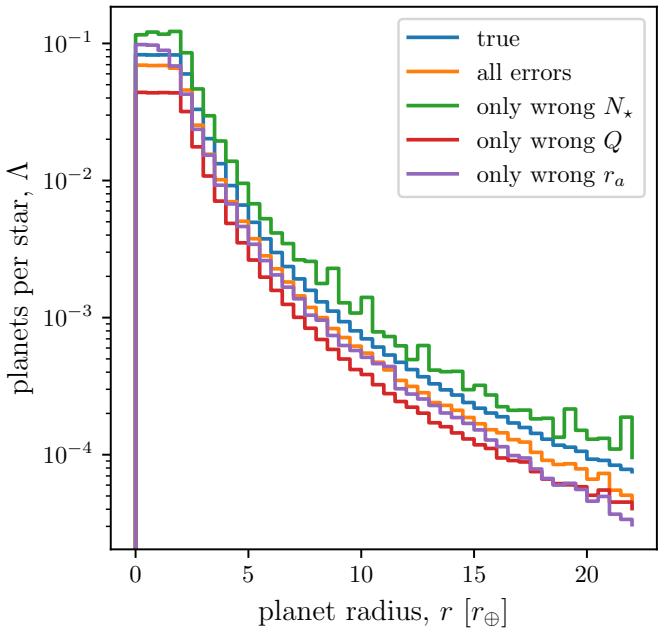


**Figure 5.** Inferred planet occurrence rates as a function of planet radius in Model #3. This model has fixed primaries and single stars, but varying secondaries. The true planet radius distribution is a power law with exponent $-2.92$ above $2R_\oplus$, below which it is uniform (e.g., Howard et al., 2012).

### 3.3. *Model #3: Fixed primaries, varying planets and secondaries*

In this model, as in the previous one, all single and primary stars have identical properties. Only the secondaries have masses, radii, and luminosities that vary between systems: $M \propto R \propto L^{1/\alpha}$. The radii of planets are assigned independently of any host system property, and are sampled from an intrinsic radius distribution, which we take as

$$p_r(r) \propto \begin{cases} r^\delta & \text{for } r \geq 2r_\oplus \\ \text{constant} & \text{for } r \leq 2r_\oplus. \end{cases} \tag{29}$$

Following Howard et al. (2012)'s measurement, we take $\delta = -2.92$. Our "nominal model" remains the same: the binary fraction is 0.44, $\alpha = 3.5$, $\beta = \gamma = 0$. We take $Z_i$, the occurrence rate integrated over all planet radii for the $i^{\text{th}}$ system type, to be equal for singles, primaries, and secondaries.

For this model, we forgo analytic development and focus only on numerics. The occurrence rates are shown as a function of planet radius in Fig. 5. For the assumed planetary and stellar distributions, the inferred rate is underestimated over all radii.

*Hot Jupiter Occurrence Rates*—Taking Fig. 5 and counting the number of planets per star with $r > 8r_\oplus$, we can compare the true and inferred hot Jupiter occurrence rates. Under the above assumptions, the true rate is 9.1 hot Jupiters per thousand stars. The inferred rate is 6.9 per thousand stars. This means that the inferred rate underestimates the true rate by a multiplicative factor of $\sim$1.3.

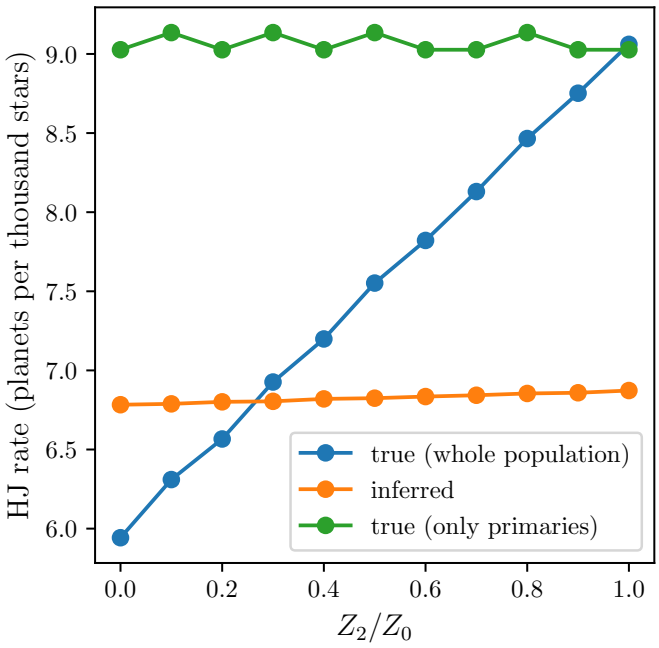


**Figure 6.** $X_i$ is the occurrence rate integrated over all possible phase space for the $i^{\text{th}}$ system type. $Z_2/Z_0 = 1$ corresponds to an equal number of planets per secondary as per single star; $Z_2/Z_0 = 0$ corresponds to secondaries not having any planets. In our Model #3, though the true HJ occurrence rate is highly dependent on $Z_2$, the inferred rate hardly depends on whether secondaries have HJs. This means that the "correction factor" between the inferred rate and the true rate around single stars is underestimated by a multiplicative factor of $\approx 1.3$, independent of the HJ rate around secondaries. The "HJ rate" is the summed rate from Fig. 5 above $8r_\oplus$.

However, this result only applies under the assumption that $Z_0 = Z_1 = Z_2$. If hot Jupiters are less common around lower mass stars, it would be more sensible to consider $Z_2 < Z_0$, while letting single stars and primaries host planets at the same rate. Therefore in Fig. 6 we let $Z_2$ vary, and show the resulting inferred and true hot Jupiter ($r > 8r_\oplus$) rates. The result is that the inferred rate is nearly independent of $Z_2$ – this is because most ($< 1/10$) secondaries are not searchable, and so their completeness fraction is much smaller than that of primaries or single stars. This means far fewer detected hot Jupiters orbit secondaries, and so they hardly affect the inferred rate. While the "true rate" across the entire population is linearly dependent on $Z_2$, the rate around singles and primaries (green line) is independent of that around secondaries. Assuming that RV surveys are measuring the true rate around single stars (or primaries), this suggests that binarity might contributes to the HJ rate discrepancy at the $\sim 0.2\%$ level, independent of the HJ rate around secondaries.

*The Rate of Earth Analogs*—The rate (density) of Earth-like planets orbiting Sun-like stars has been independently measured by Youdin, Petigura, Dong & Zhu, Foreman-Mackey et al., and Burke et al., (2011, 2013, 2013, 2014, and 2015 respectively). These efforts have found that the one-year terrestrial planet occurrence rate varies

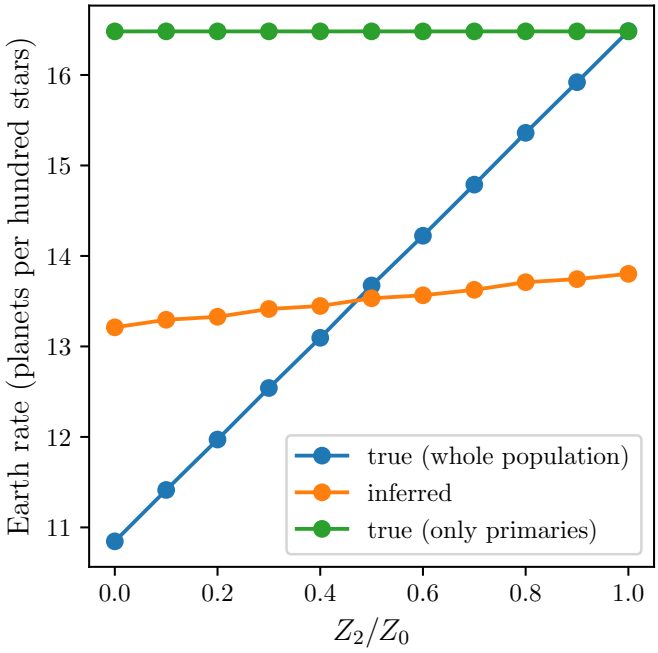


**Figure 7.** Same as Fig. 6, but for Earth-sized planets. The absolute values given on the $y$-axis found by summing the rate from Fig. 5 for planetary radii from 0.5 to $1.5r_\oplus$ (this is a toy model, and they do not reflect an actual determination of $\eta_\oplus$). The relative values show that the inferred rate for Earths is roughly independent of the occurrence rate (integrated over all radii) around secondaries. However, it is systematically lower than the true rate around single and primary stars, by $\approx 20\%$.

between $\approx 0.03$ and $\approx 1$ per Sun-like star, depending on assumptions that are made when retrieving the rate (Burke et al. 2015's Fig. 17).

Our model does not explicitly include the rate density's period-dependence, because stellar binarity does not appreciably bias the period-dependence of occurrence rates measured by transit surveys[4]. Instead, it allows us to evaluate the difference in the apparent and true rate as a function of radius (Fig. 5). At Earth's radius, the result is that the inferred rate is $0.84\times$ the true rate around single stars, assuming that the $Z_i$'s are equal. Similar to the above case of the hot Jupiters, if we vary the true $Z_2$ while keeping $Z_0 = Z_1$, the ratio of the inferred to true rate around single stars hardly changes by only a few percent (shown in Fig. 7). The ratio of the inferred to the true rate, $(\Lambda_{\rm inferred}/\Lambda)_{r=r_\oplus}$ varies substantially, but by at most 50% in the (unrealistic) limiting case that secondaries do not host planets.

## 4. DISCUSSION

[4] Note that stellar binarity does bias the *intrinsic* planet occurrence as a function of planetary and binary periods. This is expected from dynamical stability limits in $\geq$3 body systems, and has been observationally confirmed (theory by Holman & Wiegert 1999, and others including Gongjie; confirmation from Wang et al 2014a, 2014b, Kraus et al 2014). However our statement is that inferred rates as a function of planet period should be negligibly affected by this, given the geometric bias against long-period transit detections, and the fact that the period distribution of solar-type binaries peaks at $\approx 100$ years (Raghavan et al 2010, Fig. 13).

*How Has Binarity Been Considered in Occurrence Rate Measurements?*—Many authors have computed planet occurrence rates using transit survey data[5]. Notable studies using *Kepler* data include those by Howard et al. 2012, Fressin et al., 2013, Foreman-Mackey et al., 2014, Dressing & Charbonneau 2015, and Burke et al. 2015. Mostly if not entirely, these studies have ignored stellar multiplicity. However, binarity clearly introduces a level of systematic uncertainty to pipeline completeness, as well as to star and planet counts. No one has yet carefully studied the magnitude of this issue for *Kepler* occurrence rates. While the present work does not resolve this problem, it suggests the approximate order of magnitude for the necessary correction factors.

Of course, on a system-by-system level stellar multiplicity affects the interpretation of planet candidates. High resolution imaging campaigns have been undertaken to measure the multiplicity of almost all *Kepler* Objects of Interest (Howell et al. 2011; Adams et al. 2012, 2013; Horch et al. 2012, 2014; Lillo-Box et al. 2012, 2014; Dressing et al. 2014; Law et al. 2014; Cartier et al. 2015; Everett et al. 2015; Gilliland et al. 2015; Wang et al. 2015a, 2015b; Baranec et al. 2016). The results of these programs have been collected by Furlan et al. (2017), and they represent an important advance in understanding the KOI sample's multiplicity statistics. In particular, they can be immediately applied to rectify binarity's influence on the mass-radius diagram (Furlan & Howell 2017).

To help clarify *Kepler*-derived occurrence rates, the high resolution imaging campaigns have not yet come full circle by observing a comparison sample of non-KOI host stars. The most recent rate studies have thus used Furlan et al. (2017)'s catalog to test the effects of removing KOI hosts with known companions, which is an important step towards reducing contamination in the "numerator" of the occurrence rate (Fulton et al. 2017). However, without an understanding of the multiplicity statistics of the non-KOI hosting stars that are assumed to be searchable, the true completeness, the true number of searchable stars, and thus the true occurrence rates will remain uncertain.

*The Hot Jupiter Rate Discrepancy*—There is at least one context in which measurement of occurrence rates may already be showing the signatures of binarity. Hot Jupiter occurrence rates measured by transit surveys ($\approx 0.5\%$) are marginally lower than those found by radial velocity surveys ($\approx 1\%$; see Table 1). The discrepancy has weak statistical significance ($< 3\sigma$). That said, one reason to expect a difference is that the corresponding stellar populations have distinct metallicities. As argued by Gould et al. (2006), the RV sample is biased towards metal-rich stars, which have been measured by RV surveys to preferentially host more giant planets (Santos et al 2004, Fischer and Valenti 2005). The *Kepler* sample specifically has been measured to be more metal poor than the local neighborhood, with a mean metallicity of

[5] An online list of occurrence rate papers is maintained at https://exoplanetarchive.ipac.caltech.edu/docs/occurrence_rate_papers.html

$[\mathrm{M/H}]_{\mathrm{mean}} \approx -0.05$ (Dong et al., 2014; Guo et al., 2017). Studying the problem in detail, Guo et al. recently argued that the metallicity difference could account for a $\approx 0.1\%$ difference in the measured rates between the CKS and *Kepler* samples – not a $\approx 0.5\%$ difference. Guo et al. concluded that "other factors, such as binary contamination and imperfect stellar properties" must also be at play.

Aside from surveying stars of varying metallicities, radial velocity and transit surveys differ in how they treat binarity. Radial velocity surveys typically reject both visual and spectroscopic binaries (*e.g.*, Wright et al. 2012). Transit surveys typically observe binaries, but the question of whether they were searchable to begin with is usually left for later interpretation. In spectroscopic follow-up of candidate transiting planets, the prevalence of astrophysical false-positives may also lead to a bias against confirmation of transiting planets in binary systems.

Ignoring these complications, in this work we showed that binarity does bias transit survey occurrence rates, simply through its effects on the number of searchable stars and the apparent radii of detected planets. Specifically, our results from Sec. 3.3 indicate that binarity could lead to underestimated HJ rates by a multiplicative factor of $\approx 1.3$.

We assess the effect this might have towards resolving the hot Jupiter rate discrepancy by asking: what is the probability of Wright et al. (2012)'s result, given a rate drawn from the stated bounds of Petigura et al. (in prep)? (See Table 1). In other words, we first take the true HJ rate as $\Lambda_{\mathrm{HJ}} = 5.7 \pm 1.3$, with Gaussian uncertainties, and then drawing from a Poisson distribution compute the probability of detecting at least 10 hot Jupiters in a sample of 836 stars. Without accounting for binarity or metallicity, only 4% of RV surveys would detect at least 10 hot Jupiters. If we multiply $\Lambda_{\mathrm{HJ}}$ by 1.2 to account for Guo et al. (2017)'s measured metallicity difference between the *Kepler* field and the local solar neighborhood, 9% of RV surveys would detect at least 10 hot Jupiters. If we then multiply again by 1.3 to account for binarity's bias, we find that 23% of RV surveys would detect at least 10 hot Jupiters, and any discrepancy would be quite tenuous We emphasize that this result is only suggestive – a true resolution of the rate discrepancy would likely require a detailed understanding of the *Kepler* field's multiplicity statistics.

*Does a detected planet orbit the primary or secondary?*—Ciardi et al. (2015) studied the effects of stellar multiplicity on the planet radii derived from transit surveys. They modeled the problem for *Kepler* objects of interest by matching a population of binary and tertiary companions to KOI stars, under the assumption that the KIC-listed stars were the primaries. They then computed planet radius correction factors assuming that *Kepler*-detected planets orbited the primary or companion stars with equal probability (their Sec. 5). Under these assumptions, they found that any given planet's radius is on average underestimated by a multiplicative factor of 1.5.

Our models show that assuming a detected planet has equal probability of orbiting the primary or secondary leads to an overstatement of binarity's population-level

effects. A planet orbiting the secondary does lead to extreme corrections, but these cases are rare outliers, because the searchable volume for secondaries is so much smaller than that for primaries. Phrased in terms of the completeness, in our Model #3 only $\sim 6\%$ of selected secondaries are searchable, compared to $\sim 60\%$ of selected primaries. This means that when high-resolution imaging discovers a binary companion in system that hosts a detected transiting planet, the planet is much more likely to orbit the primary. This statement is independent of the fact that planets are often confirmed to orbit the primary by inferring the stellar density from the transit duration.

*On the utility for future occurrence rate measurements*—Though they will be difficult to distinguish from false positives, *TESS* is expected to discover over $10^4$ giant planets (Sullivan et al. 2015). One possible use of this overwhelmingly large sample will to measure an occurrence rate of short-period giant planets. Our work indicates that if this measurement is to be more precise than $\sim 30\%$, binarity cannot be neglected.

*What about Kepler?*—Barclay & Collaborator (in preparation) have performed the exercise of taking stars selected by the *Kepler* team, pairing them with a population of secondaries, injecting a realistic distribution of planet radii, and then comparing the inferred occurrence rates with the true ones. In their model, they find that binarity leads to an inferred rate of Earth-sized planets $\approx 10\%$ less than the true rate. In our Model #3, if all $Z_i$'s are equal (a plausible assumption in the lack of evidence to the contrary), the underestimate is by 16%.

## 5. CONCLUSION

This study presented three simple models for the effects of binarity on occurrence rates measured by transit surveys. The most realistic of these models (Model #3) suggests that binarity does lead to underestimates in transit survey occurrence rates, but with less than 30% relative error. The model further suggests that hot Jupiter rates measured by transit surveys are biased to infer $\approx 1.3\times$ fewer hot Jupiters per star than surveys that only measure occurrence rates about single stars (*i.e..,* radial velocity surveys). It also indicates that binarity's effects on the measured occurrence rates of Earth-sized planets are far smaller than current systematic uncertainties. Though our models are simplistic, their agreement with Barclay & Collaborator's recent detailed simulations indicate that they may capture the essential ingredients.

**Table 1.** Occurrence rates of hot Jupiters (HJs) about FGK dwarfs, as measured by radial velocity and transit surveys.

| Reference | HJs per thousand stars | HJ Definition |
|---|---|---|
| Marcy+ 2005 | 12±1 | $a < 0.1\,\mathrm{AU}; P \lesssim 10\,\mathrm{day}$ |
| Cumming+ 2008 | 15±6 | – |
| Mayor+ 2011 | 8.9±3.6 | – |
| Wright+ 2012 | 12.0±3.8 | – |
| Gould+ 2006 | $3.1^{+4.3}_{-1.8}$ | $P < 5\,\mathrm{day}$ |
| Bayliss+ 2011 | $10^{+27}_{-8}$ | $P < 10\,\mathrm{day}$ |
| Howard+ 2012 | 4±1 | $P < 10\,\mathrm{day}; r_p = 8 - 32 r_\oplus$; solar subset[a] |
| – | 5±1 | solar subset extended to $Kp < 16$ |
| – | 7.6±1.3 | solar subset extended to $r_p > 5.6 r_\oplus$. |
| Moutou+ 2013 | 10±3 | *CoRoT* average; $P \lesssim 10\,\mathrm{day}$, $r_p > 4 r_\oplus$ |
| Petigura+ (in prep) | 5.7±1.3 | $r_p = 8 - 24 r_\oplus$; $1 < P/\mathrm{day} < 10$; CKS stars[b] |
| Santerne+ (in prep) | 9.5±2.6 | *CoRoT* galactic center |
| – | 11.2±3.1 | *CoRoT* anti-center |

NOTE— The upper four results are from radial velocity surveys; the rest are from transit surveys. Many of these surveys selected different stellar samples. "–" denotes "same as above".

[a] Howard+ 2012's "solar subset" was defined as *Kepler*-observed stars with $4100\,\mathrm{K} < T_{\mathrm{eff}} < 6100\,\mathrm{K}$, $Kp < 15$, $4.0 < \log g < 4.9$. They required signal to noise > 10 for planet detection.

[b] Petigura+ (in prep)'s CKS stars are, for the most part, all KOIs with $Kp < 14$. **the stars assumed to be searchable were**