SEARCHING FOR COMETS ON THE WORLD WIDE WEB: THE ORBIT OF 17P/HOLMES FROM THE BEHAVIOR OF PHOTOGRAPHERS

DUSTIN LANG\textsuperscript{1} AND DAVID W. HOGG\textsuperscript{2,3}

\textsuperscript{1} Princeton University Observatory, Princeton, NJ 08544, USA; dstn@astro.princeton.edu
\textsuperscript{2} Center for Cosmology and Particle Physics, Department of Physics, New York University, 4 Washington Place, New York, NY 10003, USA
\textsuperscript{3} Max-Planck-Institut für Astronomie, Königstuhl 17, D-69117 Heidelberg, Germany

ABSTRACT

We performed an image search for “Comet Holmes,” using the Yahoo! Web search engine, on 2010 April 1. Thousands of images were returned. We astrometrically calibrated—and therefore vetted—the images using the Astrometry.net system. The calibrated image pointings form a set of data points to which we can fit a test-particle orbit in the solar system, marginalizing over image dates and detecting outliers. The approach is Bayesian and the model is, in essence, a model of how comet astrophotographers point their instruments. In this work, we do not measure the position of the comet within each image, but rather use the celestial position of the whole image to infer the orbit. We find very strong probabilistic constraints on the orbit, although slightly off the Jet Propulsion Lab ephemeris, probably due to limitations of our model. Hyperparameters of the model constrain the reliability of date meta-data and where in the image astrophotographers place the comet; we find that ~70% of the meta-data are correct and that the comet typically appears in the central third of the image footprint. This project demonstrates that discoveries and measurements can be made using data of extreme heterogeneity and unknown provenance. As the size and diversity of astronomical data sets continues to grow, approaches like ours will become more essential. This project also demonstrates that the Web is an enormous repository of astronomical information, and that if an object has been given a name and photographed thousands of times by observers who post their images on the Web, we can (re-)discover it and infer its dynamical properties.

Key words: celestial mechanics – comets: individual (17P/Holmes) – ephemerides – methods: statistical – surveys – time

Online-only material: color figures

1. INTRODUCTION

The Web bristles with billions of images: on Web sites, in public photo-sharing sites, on social networks, and in private email and file-sharing conversations.\textsuperscript{4} A tiny fraction but enormous number of these images are astronomical images—images of the night sky in which astronomical sources are visible. This is true even if we exclude from consideration scientific collections such as those of professional observatories and surveys and only count the images of hobbyists, amateurs, and sightseers. In principle, these images, taken together, contain an enormous amount of information about the astronomical sky. Of course they have no scientifically responsible provenance, have never been “calibrated” in any sense of that word, and were (mainly) taken for purposes that are not at all scientific, but having been generated from CCD-like measurements of the intensity field, they cannot help but contain important scientific information. The Web is, therefore, an enormous and virtually unexploited sky survey.

It is difficult to estimate the total number of astronomical images on the Web, and even harder to estimate the total data throughput (\textit{étendue} or equivalent measure of scientific information content). However, by any estimate, it is extremely large. For example, image search results for common astronomical subjects include thousands of astronomical images. The Flickr photo-sharing Web site has an astrometry group (administered by the Astrometry.net collaboration; more below) with more than 23,000 photos, and its astronomy and astrophotography groups have more than 35,000 and 45,000, respectively. A search for the Orion Nebula on Flickr returns more than 9000 images, which jointly contain significant information on very faint stars and nebular features. These numbers—derived solely from Flickr searches—represent only a tiny fraction of the relevant Web content. Of course all these search results contain many nonastronomical images, diagrams, fake data, and duplicates, so use of them for science is nontrivial.

The technical obstacles to making use of Web data are immense: if anything has been learned from our interaction with online communication, it is that publisher-supplied or provider-supplied meta-data about Web content are consistently missing, misleading, in error, or obscure. Indeed, when it comes to the astronomical properties of imaging discovered on the Web, most providers do not even know what we want in terms of “meta-data;” we want calibration parameters relating to image date, astrometric coordinate system, photometric sensitivity, and point-spread function, and we want it in machine-readable form. The Virtual Astronomy Multimedia Project (Gauthier et al. 2008) has defined a format for placing astrometric meta-data in image headers, but the goal of the project is to make “pretty pictures” searchable for education and public outreach purposes, rather than science. They do not consider the problem of producing or verifying calibration meta-data; they assume correct meta-data are provided along with the science images that are used to produce the pretty pictures. Even the Virtual Observatory (http://ivoa.net), which concentrates on astronomical meta-data, has no plan for ensuring that meta-data are correct, and has no machine-readable form for...
many quantities of great interest (such as the detailed point-spread-function model); we cannot expect the world’s amateur astrophotographers to be better organized.

Two important changes are occurring in astronomy that are opening up the possibility that we might exploit data collections as radically confusing as that of the entire Web. The first is that tools are beginning to appear that can perform completely hands-free data analysis tasks. The best example so far is the Astrometry.net system, which can take astronomical imaging of completely unknown provenance, and calibrate it astrometrically using the data in the image pixels alone (Lang et al. 2010).

The second change is that there has been an enormous increase in the amount and diversity of publicly available professional data—that is, calibrated, trustworthy, science-oriented data in observatory, sky-survey, and individual-investigator collections. These collections are so large and diverse that automated data analysis tools that can trivially interact with extremely heterogeneous data are necessary in many scientific domains. That is, much of the technology required for exploitation of the Web-as-sky-survey is required for any mature, data-intensive scientific investigation.

We have been exploring some of these ideas with the Astrometry.net project. Not only has the system calibrated thousands of images taken by amateurs and hobbyists, we have interfaced the system with Flickr (Stumm et al., forthcoming). When a user adds an image to the “astrometry” group, an automated “bot” downloads the image, calibrates it with Astrometry.net, and then posts machine-readable calibration results on the image’s page on Flickr; these have been dubbed “astro-tags.” The bot also adds annotations to the image, marking named stars and galaxies from the Messier and NGC/IC catalogs. We make use of the Flickr Application Programming Interface (API); many image and data-sharing sites offer APIs that allow scriptable access to the data they hold. The success of our Flickr bot suggests that automated maintenance of a heterogeneous crowd-sourced sky survey might be possible in the future.

In this paper, we explore some of the ideas around the Web-as-sky-survey, by performing a scientific investigation of Comet 17P/Holmes using Web-discovered, human-viewable (JPEG) images alone. Although we do in principle learn things about Comet Holmes, our main interest is in developing and testing automated data analysis tools that can trivially interact with the data set we use in our analysis below. Figure 2 shows the footprints of the images on the sky.

Many of the images in this data set are annotated images or diagrams such as finding charts or illustrations of the comet’s orbit. Some of these diagrams were recognized by Astrometry.net as images of the sky. This can be seen in the co-added image in Figure 2, where there are clearly lines connecting the stars that form the constellation Perseus.

Of the 1299 images in our data set, 422 have timestamps in the image headers (“Exchangeable image file format” or EXIF headers). The distribution of timestamps is shown in Figure 3. On 2007 October 24, Comet 17P/Holmes brightened by more than 10 mag (Buzzi et al. 2007), generating considerable public interest and making it a very popular and accessible observing target in the amateur astronomy community. The distribution of image timestamps shows a large spike at this time.

We evaluate the accuracy of the image timestamps by asking, for each image, whether the comet would appear within the celestial-coordinate bounds of the image at its stamped time. We find that the majority of the timestamps are consistent, and that inconsistent timestamps are typically late rather than early; we assume this is because some of the images have been post-processed and the timestamp represents the time the image was last edited rather than the time the image was taken. See Figure 4.

Figure 3 shows the distribution of angular scales of the images in our data set. The distribution peaks around 3 deg2. Also shown is the distribution of exposure times reported in the EXIF headers.

3. ORBIT INFERENCE

We take the approach of generative modeling; that is, we construct a well-defined approximation to the probability of the data given the model. We take the “data” to be the pointing (on the sky) of each astrometrically calibrated image (as determined by Astrometry.net); recall that the goal is to use the behavior of astrophotographers (in pointing their cameras) to find the gravitational orbits of objects in the sky. We treat the time at which each image is taken as a hidden “nuisance” parameter.

For each image $i$ there is a pointing $\alpha_i$ (two-dimensional position or celestial coordinates on the sky). These are the data. The image was taken at time $t_i$, and has image parameters $\Omega_i$ (camera plate scale, image size, orientation, and reported EXIF timestamp if there is one), taken to be known. In addition, the comet has orbital parameters $\omega$, which can be thought of

2. DATA AND CALIBRATION

Our data collection began with a search of the World Wide Web. We used the pYsearch (Hedstrom 2007) code to access the Yahoo! Web Search service. On 2010 April 1, we searched for JPEG-format images using the query phrase “Comet Holmes.” Due to a dramatic brightening during its 2007 apparition, Comet Holmes became a very popular and accessible target for astrophotographers, so many images are available on the Web. Our search yielded approximately 10,000 total results, but the Yahoo! Web Search API allowed only 1000 results to be retrieved per query. In order to broaden the result set, we performed an additional set of searches: for each Web site containing an image in the original set of results, we performed a query that was limited to that Web site. These queries, performed later on 2010 April 16, produced an additional 2741 results (including some duplicates), for a total of 2476 unique results. See Figure 1 for some example images.

Next, we retrieved the images on 2010 April 16. This yielded a total of 2309 valid JPEG images. After removing byte-identical images, 2241 unique images remained. We then ran the Astrometry.net code on each image to perform astrometric calibration. A total of 1299 images were recognized as images of the night sky and astrometrically calibrated. These images form the data set we use in our analysis below. Figure 2 shows the footprints of the images on the sky.

On 2007 October 24, Comet 17P/Holmes brightened by more than 10 mag (Buzzi et al. 2007), generating considerable public interest and making it a very popular and accessible observing target in the amateur astronomy community. The distribution of image timestamps shows a large spike at this time.

We evaluate the accuracy of the image timestamps by asking, for each image, whether the comet would appear within the celestial-coordinate bounds of the image at its stamped time. We find that the majority of the timestamps are consistent, and that inconsistent timestamps are typically late rather than early; we assume this is because some of the images have been post-processed and the timestamp represents the time the image was last edited rather than the time the image was taken. See Figure 4.

Figure 3 shows the distribution of angular scales of the images in our data set. The distribution peaks around 3 deg2. Also shown is the distribution of exposure times reported in the EXIF headers.

3. ORBIT INFERENCE

We take the approach of generative modeling; that is, we construct a well-defined approximation to the probability of the data given the model. We take the “data” to be the pointing (on the sky) of each astrometrically calibrated image (as determined by Astrometry.net); recall that the goal is to use the behavior of astrophotographers (in pointing their cameras) to find the gravitational orbits of objects in the sky. We treat the time at which each image is taken as a hidden “nuisance” parameter.

For each image $i$ there is a pointing $\alpha_i$ (two-dimensional position or celestial coordinates on the sky). These are the data. The image was taken at time $t_i$, and has image parameters $\Omega_i$ (camera plate scale, image size, orientation, and reported EXIF timestamp if there is one), taken to be known. In addition, the comet has orbital parameters $\omega$, which can be thought of
Figure 1. Example images from the image search. Here we have highlighted the diversity of the images; the majority are in fact high-quality, narrow-field images of the comet, such as (b), (h), (m), and (u). Images that did not calibrate successfully with Astrometry.net (and therefore were not used as data in this study) are marked with asterisks. Note that images (e), (p), and (q) were all successfully calibrated and were used in the analysis. Image (a) shows a statue of Perseus, a constellation through which Comet Holmes traveled during its 2007 approach. Image (d) is from the year 1892, shortly after the discovery of Comet Holmes. Although it was calibrated and used in our analysis, this image was identified as an outlier since we are fitting only the 2007 apparition of the comet. Image (f) includes the California nebula (NGC 1499). Credits: (a) copyright 2000–2005 Gods, Heroes, and Myth (http://www.gods-heros-myth.com); (b) Paolo Berardi; (c) Amateur Astronomers, Inc. Research Committee (http://asterism.org); (d) Edward Emerson Barnard (Ball 1905); (e) copyright P.-M. Hedén (http://www.clearskies.se); (f) copyright Dave Kodama (http://astrocamera.net); (g) TOC Observatory (http://tocobs.org); (h) copyright Fay Saunders; (i) Bruce Card, Aldrich Astronomical Society, Worcester MA; (j) NASA, JPL-Caltech, W. Reach (SSC-Caltech); (k) copyright Julián Cantarelli; (l) copyright 2007, 2008 John F. Pane (http://holmes.johnpane.com); (m) copyright Tyler Allred (http://allred-astro.com); (n) Joe Orman; (o) NASA, ESA, and H. Weaver (The Johns Hopkins University Applied Physics Laboratory), and A. Dyer, Alberta, Canada; (p) copyright Vincent Jacques (http://vjac.free.fr/skyshows); (q) Jimmy Westlake, Colorado Mountain College; (r) Stephane Zoll (http://astrosurf.com/zoll); (s) Babak Tafreshi / TWAN (http://twanight.org); (t) Ivan Eder (http://eder.csillagaszat.hu); (u) Vicent Peris (OAUUV), José Luis Lamadrid (CEFCA); (v) copyright Thorsten Boeckel (http://boeckel.de); (w) D. J. Barry, Department of Astronomy, Cornell University.
as semimajor axis, eccentricity, inclination, longitudes, etc., or equivalently a three-dimensional position $x$ and velocity $v$ at a chosen epoch. We choose the latter for inference simplicity, and use as the epoch JD 2454416.0 (2007 November 12).

Finally, there are three additional nuisance hyperparameters $\theta$ that will appear as we go. In this work we consider only the 2007 apparition of the comet. Our data set does include at least one image of the comet during an earlier apparition (1892; see Figure 1(d)), but we chose not to attempt to fit multiple apparitions.

The single-image likelihood $p(\alpha_i|\Omega_i, \omega, \theta)$ is a marginalization over time $t_i$ of the time-dependent single-image likelihood,

$$p(\alpha_i|t_i, \Omega_i, \omega, \theta):$$

$$p(\alpha_i|\Omega_i, \omega, \theta) = \int p(\alpha_i|t_i, \Omega_i, \omega, \theta) p(t_i|\Omega_i, \omega, \theta) dt_i,$$  \hspace{1cm} (1)

where the time-dependent single-image likelihood is a mixture of “foreground” (inlier) and “background” (outlier) components:

$$p(\alpha_i|t_i, \Omega_i, \omega, \theta) = p_{\text{good}} p_{\text{fg}}(\alpha_i|t_i, \Omega_i, \omega, \theta)$$

$$+ [1 - p_{\text{good}}] p_{\text{bg}}(\alpha_i).$$  \hspace{1cm} (2)
Figure 3. Top left: the timestamps in the EXIF headers of the images in our data set. Of the 1299 images 422 have timestamps; 394 are within the range plotted here. The dashed line marks 2007 October 24, the date of the comet’s outburst and dramatic brightening (Buzzi et al. 2007). Top right: the distribution of camera manufacturers listed in the EXIF headers (for the 183 images containing manufacturer information). Bottom left: the distribution of image angular sizes in our data set, according to Astrometry.net. Bottom right: the distribution of exposure times in our data set, according to EXIF headers (for the 149 images containing an exposure time entry).

(A color version of this figure is available in the online journal.)

The components are

\[
p_{fg}(\alpha_i|\Omega_i, \omega, \theta) = \begin{cases} 
[\eta \Omega_i]^{-1} & \text{comet in } \eta \text{ sub-image} \\
0 & \text{comet not in } \eta \text{ sub-image}
\end{cases}
\]

(3)

\[
p_{bg}(\alpha_i) = [4\pi]^{-1}.
\]

(4)

The hyperparameters \(\theta\) include \(p_{\text{good}}, \eta,\) and \(p_{\text{EXIF}}\) (discussed below). Here, \(p_{\text{good}}\) is the probability that the image really is a picture intentionally taken of (generated by) the comet, \(p_{fg}(\cdot)\) is a “foreground” model, which gives high likelihood when the comet (with orbital parameters \(\omega\) at time \(t_i\)) is inside the image, and \(p_{bg}(\cdot)\) is a “background” model, with no dependence on the comet or time, that describes images that are in our data set but do not contain the comet (perhaps because they were incorrectly returned by the Web search engine). The \(4\pi\) in \(p_{bg}(\cdot)\) is the solid angle of the whole sky. The hyperparameter \(\eta\) (subject to \(0 < \eta < 1\)) controls the fractional size of the central region of an image in which astrophotographers place comet subjects, \(\Omega_i\) is the solid angle covered by image \(i\), and the “\(\eta\) subimage” is the central \(\eta\) of the image. In detail, we define the \(\eta\) sub-image to have the same aspect ratio as the whole image, centered at the same point, but smaller in angular size by \(\sqrt{\eta}\) along both dimensions.

Our Bayesian prior probability distribution function (PDF) over time, \(p(t_i|\Omega_i, \theta)\), turns out to be crucial to good inference in this problem, in part because trivial or wrongly uninformative time PDFs lead to highly biased answers, a point to which we will return below. We expect that a large fraction of the image EXIF timestamps (where they exist) are correct, but at the same time we cannot trust them completely. We construct an empirical “cheater” prior \(p_{\text{emp}}(t)\) based on the empirical
Figure 4. Evaluation of the accuracy of the timestamps in the image EXIF headers. For each image, we computed the range of times that the comet appeared inside the celestial-coordinate bounds of the image; that is the shaded gray region, which is vertically centered at zero. The images are sorted so that this envelope is monotonic. The EXIF timestamp for each image is shown as a bar of height one day (since EXIF has no time zone specification, this is our intrinsic uncertainty), plus a minimum size to make all the markers visible. If the EXIF timestamps were correct and set to UT, the position of the bars touch the gray region, indicating that the majority of the EXIF timestamps were correct and set to UT, the position of the bar within the gray region would indicate the position of the comet within the image. Most of the timestamps were correct and set to UT, the position of the bar within the gray region, indicating that the majority of the EXIF timestamps were correct and set to UT, the position of the bar within the gray region would indicate the position of the comet within the image. Most of the timestamps were correct and set to UT, the position of the bar within the gray region would indicate the position of the comet within the image. Most of the timestamps were correct and set to UT, the position of the bar within the gray region would indicate the position of the comet within the image.

Figure 5. Time prior PDF $p_{\text{emp}}(t)$ used for images with no EXIF date information (dark line) and the time prior PDF for an image with a particular EXIF date (lighter line). The latter is a mixture of $p_{\text{emp}}(t)$ and a top-hat of width one day centered on the EXIF date and fractional weight $p_{\text{EXIF}} = 0.71$. We use a width of one day because the EXIF standard does not permit the time zone to be specified. The function $p_{\text{emp}}(t)$ is based on the distribution of reported EXIF dates shown in Figure 3; details in the text.

The single-image likelihood $p(\alpha | \Omega, \omega, \theta)$ is the probability that an image would be taken at coordinates $\alpha$, given the camera properties $\Omega$, the comet’s trajectory $\omega$, and our model hyperparameters $\theta$, integrated (marginalized) over the time period we consider. We do not specify the exact time $t_i$ of each image; we instead specify a probability distribution of times and integrate over it (and this integration requires us to be Bayesian). In order to evaluate this likelihood, we compute the comet trajectory on a fine time grid and perform the time integral numerically as a sum over grid points. For dynamical integration we use a Keplerian two-body celestial mechanics code implemented in Python by Astrometry.net for both the comet and the Earth–Moon barycenter (EMB); we take the initial conditions of the EMB from JD 2454101.5 (2007 January 1). For simplicity we take the EMB to be the observer’s location, thus ignoring the effect of parallax. At the precision of the data, the finite light-travel time in the solar system is significant; we include it when we consider the observed position of the comet as a function of time. For the numerical integrals, we simply convert observed solar system directions to positions on the celestial sphere, and positions on the celestial sphere to image positions (to determine whether particular comet instances are inside particular images) with the Astrometry.net world-coordinate system libraries (Lang et al. 2010). Outliers—images that do not contain the comet—will only by chance intersect the comet trajectory so will have low likelihood under the foreground model $p_B(\cdot)$; the background model $p_{\text{bg}}(\cdot)$, which has no dependence on the orbital parameters, will dominate the likelihood. As with the time $t_i$, we need not explicitly estimate whether any given image is an outlier; we instead model the data as a mixture of the foreground
distribution ("inliers") and a background distribution ("outliers"), and sum (marginalize) over the two possibilities.

The total likelihood is the product of the individual-image marginalized likelihoods, and the posterior PDF for the parameters \(\omega\) and \(\theta\) is proportional to the total likelihood times a prior. We take this prior to be Gaussian in comet position \(x\) with three-dimensional isotropic Gaussian variance of \(1\) AU\(^2\), a beta distribution in squared velocity \(v^2 \equiv v \cdot v\) between \(v^2 = 0\) and the \(v^2\) that just unbinds the comet, with beta-distribution parameters \(\alpha = 1\) and \(\beta = 3\). We take the prior to be flat in the range 0–1 for the probability hyperparameters \(p_{\text{good}}\) and \(p_{\text{EXIF}}\) and (improperly) flat in \(\ln(\eta)\) for the fractional hyperparameter \(\eta\). These nine parameters (three position components, three velocity components, two probabilities, and one fraction) are the parameters in which we perform our Markov Chain Monte Carlo (MCMC) sampling. We perform the sampling with a Python implementation (Foreman-Mackey et al. 2012) of an affine-invariant ensemble sampler (Goodman et al. 2010; Hou et al. 2012) using an ensemble of 64 walkers and Python multiprocessor support.

We initialize the MCMC using the following heuristic. Since we know we have many images taken during a short window around the comet’s outburst, we select images with timestamps within a week of the median date. Of these images, we keep those whose centers are within 5° of the median right ascension and declination. This gives us a set of 136 images to which we fit lines for right ascension and declination with respect to time, weighting by the inverse image extents. Assuming the comet is 1 AU away from Earth and moving perpendicular to the line of sight, we convert the projected position to a three-dimensional position \(x_0\) and velocity \(v_0\) at the epoch. We choose eyeballed-sensible initial values for the hyperparameters: \(p_{\text{good}} = 0.85\), \(p_{\text{EXIF}} = 0.75\), and \(\eta_0 = 0.4\). To initialize the ensemble of MCMC samples, we add Gaussian noise with standard deviation \(10^{-5}\) AU to the position \(x_0\), \(10^{-3}\) AU/yr to the velocity \(v_0\), and \(10^{-2}\) to each of \(p_{\text{good}}, p_{\text{EXIF}},\) and \(\eta_0\). The details of the initialization are not critical; the MCMC sampler will explore the parameter space and find a good solution in a reasonable number of iterations as long as we initialize it somewhat near the solution.

There are several substantial limitations to this model: the prior does not even come close to representing our true prior knowledge about comets, particularly ones that are observed by photographers and posted to the Web. The total likelihood (being a product of individual-image likelihoods) assumes all the data are independent, but in reality some of the images found by the Web search are repeats, duplicates, or derived images from others. Most importantly, we make no attempt to find the comet in the image. This is a model of how astronomers point their cameras, not of the visible comet itself.

The results of the inference are shown in Figure 6 as a set of sample trajectories drawn from the Markov chain. These samples are effectively drawn from the posterior PDF marginalized over the hyperparameters. The small dispersion among the samples show that the data—just the pointings of a set of heterogeneous images—are incredibly informative about the comet orbit. In Figure 7 we show our estimates for the standard orbital elements, along with the values from the Horizons system from the Jet Propulsion Lab (JPL; Giorgini et al. 1996), which we take to be authoritative. Most of our parameter estimates are a few standard deviations away from the JPL values. Figure 8 shows that the three-dimensional orbit we infer is quite close to the JPL orbit in the regions where we have data, and the variance in our inferred orbit correctly increases in the regions where we lack data.

4. DISCUSSION

We have shown that if a solar system body has been named and hundreds of astrophotographers around the world have deliberately photographed it, we can recover its dynamical properties by a Web search operation followed by a large amount of computation. All the inference is done on image positions; we never look at the content of the images. This effectively makes the model a model of astrophotographers, because the image pointings are a record of where human observers pointed their telescopes and cameras. The six dynamical parameters are parameters of the comet to be sure, but the three hyperparameters are affected by human behavior as well as by properties of the comet. The probability \(p_{\text{good}}\) relates to the purity of image search on the Web, the probability \(p_{\text{EXIF}}\) relates to the reliability of astrophotographers’ Web-published image meta-data, and the fraction \(\eta\) relates to how astrophotographers frame their images. We find hyperparameter \(p_{\text{good}} \sim 0.9\), which indicates that a large majority of the images returned by the Web search and recognized by Astrometry.net do indeed contain Comet Holmes. We find \(p_{\text{EXIF}} \sim 0.7\), indicating that when images have timestamps in their headers, they are often correct. We find \(\eta \sim 0.36\), indicating that astrophotographers tend to place Comet Holmes in the middle third of the image area.

We expect that these hyperparameters will vary for different Web search engines, query phrases, and comets. Comets with distinctive names are likely to be better indexed by search engines, faint comets are likely to be photographed by different populations of observers, and comets with long tails are likely to be framed differently in photographs. We expect that searching for the phrase “17P/Holmes” rather than “Comet Holmes,” for
example, might result in images from more technical astronomy Web sites rather than the popular press. As a result, we might find that a higher fraction of the images contain the comet, leading to a higher inferred \( p_{\text{good}} \); we might find that the images have different distributions of angular extents and different framing, resulting in different \( \eta \), and we might expect that more of the images were captured with telescope-mounted CCDs, saved in non-JPEG formats and subsequently processed to produce the pretty picture, perhaps resulting in different timestamp properties and therefore \( p_{\text{EXIF}} \).

If we omit \( \eta \) from our model (i.e., assume it is unity), we effectively ignore the fact that astrophotographers frame their subjects, and this in turn increases the positional uncertainty of the comet, which results in poorer constraints on the orbit of the comet. By including \( \eta \), we give our model the freedom to learn how astrophotographers compose their images. The model prefers to set \( \eta \) to a rather small fraction, indicating that this freedom is useful. Including \( \eta \) means that if the comet appears in the corner of an image, the image will be treated as an outlier and will not contribute to constraining the orbit. However, this loss of constraining power is balanced by stronger constraints from all images in which the comet appears near the center. In Figure 7, there are three peaks in the histogram of \( \eta \) values. We assume the two subdominant peaks correspond to \( \eta \) values that omit one image or include one additional image in the solution, relative to the dominant peak.

The model is exceedingly crude, and the fact that our results are biased (the samples in Figure 6 are offset from the JPL trajectory) is probably in part related to this crudeness. The centering model is extremely crude; in reality there is a distribution of astrophotographers’ behavior that it ought to describe. The time model involves a hard-set empirical prior that is not justified and ought to be simultaneously optimized and marginalized out in the inference (this would be a form of hierarchical inference like in Hogg et al. 2010). The time-zone model (flat across all time zones) is also not realistic, since some
Figure 8. Three-dimensional inferred orbit (thin lines) and JPL orbit (dashed line) in heliocentric equatorial coordinates. We have plotted 32 samples from our inferred orbit distribution to show the scatter. Also shown is the JPL orbit of the Earth–Moon barycenter. Our inferred orbit captures the general shape of the trajectory of the comet and is relatively unconstrained at early and late times (when we have no data).

(A color version of this figure is available in the online journal.)

time zones are much more populated with photographers than others. Along those same lines, there is an enormous amount of external information (weather data and visibility calculations) that could further constrain the possible times and time zones. As with time zones, we do not attempt to model the positions of the astrophotographers relative to the EMB. By ignoring the resulting parallax, we incur errors of about 5″ in our estimate of the comet’s position.

Another crudeness is in the assumption of independent and identically distributed draws in the likelihood. This is not true in that some of the Web images we find are crops, edits, or diagrams made from other Web images. That is, each image is not guaranteed to be an independent datum.

Our model that astrophotographers tend to place their subjects in the center of their images ignores the fact that conjunctions of astronomical objects on the sky are often targets of interest. We suspect that we see this effect in our data set. When Comet Holmes passed near the California Nebula (NGC 1499), many photographers captured the conjunction. In Figure 6 this overdensity of images can be seen near the marked location of NGC 1499. The nebula appeared below the comet (at lower decl.), and it appears that our inferred comet trajectories have been pulled down as a result.

In some sense, this project is a citizen-science project, because it does science with data generated by non-scientists. However, it is very different from projects like SETI@Home (Korpela et al. 2009) because it makes use of participants’ intelligence, not just hardware. It is very different from projects like GalaxyZoo (Lintott et al. 2011) because it makes use of specialized astronomy knowledge among the participants; one must be a relatively avid astronomer to usefully contribute. It is very different from the projects of the AAVSO (http://aavso.org) or MicroFUN (Gould 2008) because the observers observed for reasons (for all we know) completely unrelated to our scientific goals. It is different from all of these projects in that the participants contributed unwittingly.

One interesting and ill-understood aspect of a citizen-science project of this type—where the participants are not aware that they are involved—relates to giving proper credit and obtaining proper permissions to use the images. We obtained permission to show the images shown in Figure 1 but we did not even attempt to get any permissions for the majority of the 2241 images we touched in the analysis. One encouraging lesson from this project is that the photographers we did contact were very supportive: not one rejected our request for permissions, typical responses expressed enthusiasm about being involved in a scientific paper, the majority asked to see the manuscript when it appears, some sent updated images or suggestions about which images to use, and a few offered details about the data analysis and processing that was performed. A less encouraging
lesson is that Web image search APIs have an uncertain future: the Yahoo! Web Search API is being decommissioned, as is the Google Image Search API.

The biggest lesson is that there is enormous information about astronomy available in uncurated nonprofessional images on the Web. We have only scratched this surface. Think how much better we could have done if we had gone into the images and actually made some attempt at detecting the comet! Figure 2 shows that there is far more information inside the images than in just the footprints. Figure 9 shows that there is a similarly informative body of images of Comet C/1996 B2 (Hyakutake). We have also noticed that there are thousands of images of the Orion Nebula on Flickr alone, and thousands more elsewhere on the Web; the joint information in this body of images (about the nebula and about time-domain activity therein) must be staggering. Perhaps this is not surprising given the large amount of telescope aperture and detector area owned by avid photographers. We have learned that we can do high-quality quantitative astrophysics with images of unknown provenance on the Web. Is it possible to build from these images a true sky survey? We expect the answer is “yes.”

We are very sad that Sam Roweis (Toronto, Google, NYU) was not here to collaborate on this project, after co-creating with us Astrometry.net; he influenced every aspect of this project. We obtained ideas, feedback, or code from Michael Blanton (NYU), Jo Bovy (NYU/IAS), Daniel Foreman-Mackey (NYU), Jonathan Goodman (NYU), Fengji Hou (NYU), Iain Murray (Edinburgh), and Christopher Stumm (Microsoft/Etsy). We offer sincere thanks to our anonymous reviewer for very detailed and thoughtful comments that spurred many improvements to the manuscript. We benefited from the activity of a very large number of astrophotographers and Web citizens, but in particular we got permission to show images from John F. Pane, Stephane
Zoll, Thorsten Boeckel, Vincent Jacques, Babek Tafreshi, Per Magnus Hedén, Jimmy Westlake, Vicent Peris, Paolo Berardi, Fay Saunders, Dave Kodama, Tyler Allred, Joe Orman, Ivan Eder, Ray Shapp, Flemming R. Ovesen, Torben Taustrup, and Bruce Card. This work was supported in part by NASA (grant NNX08AJ48G), the NSF (grant AST-0908357), and a Research Fellowship of the Humboldt Foundation. This research made use of the SAO/NASA Astrophysics Data System, the Yahoo! Web Search service, the Jet Propulsion Lab Horizons service, the Python programming language, and open-source software in the numpy, scipy, matplotlib, emcee, pYsearch, and WCSlib projects. All code and data used in this project are available from the authors upon request.

REFERENCES

Ball, R. S. 1905, A Popular Guide to the Heavens (New York: Van Nostrand)
Buzzi, L., Muler, G., Kidger, M., et al. 2007, IAU Circ., 8886, 1

Calabretta, M. R., & Greisen, E. W. 2002, A&A, 395, 1077
Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2012, arXiv:1202.3665
Gauthier, A., Christensen, L. L., Hurt, R. L., & Wyatt, R. 2008, in Communicating Astronomy with the Public, ed. L. L. Christensen, M. Zoulias, & I. Robson (Eugenides Foundation), 214
Giorgini, J. D., Yeomans, D. K., Chamberlin, A. B., et al. 1996, BAAS, 28, 1158
Goodman, J., & Weare, J. 2010, Commun. Appl. Math. Comput. Sci., 5, 65
Gould, A. 2008, in Manchester Microlensing Conference, ed. E. Kerins, S. Mao, N. Rattenbury, & L. Wyrzykowski (PoS; SISSA), 38
Hedstrom, L. 2007, pYsearch 3.0: Python APIs for Yahoo! Search Services, http://pYsearch.sourceforge.net/
Hogg, D. W., Myers, A. D., & Bovy, J. 2010, ApJ, 725, 2166
Hou, F., Goodman, J., Hogg, D. W., & Weare, J. 2012, ApJ, 745, 198
Korpela, E. J., et al. 2009, in ASP Conf. Ser. 420, Bioastronomy 2007: Molecules, Microbes and Extraterrestrial Life, ed. K. J. Meech, J. V. Keane, M. J. Mumma, J. L. Siefert, & D. J. Werthimer (San Francisco, CA: ASP), 431
Lang, D., Hogg, D. W., Mierle, K., Blanton, M., & Roweis, S. 2010, AJ, 139, 1782
Lintott, C., Schawinski, K., Bamford, S., et al. 2011, MNRAS, 410, 166