PHYSICAL REVIEW E **96**, 030101(R) (2017)

# Random matrices and the New York City subway system

Aukosh Jagannath[1,*] and Thomas Trogdon[2,†]

[1] *Department of Mathematics, University of Toronto, Toronto, Ontario, Canada M5S 2E4*

[2] *Department of Mathematics, University of California, Irvine, California 92697-3875, USA*

(Received 3 March 2017; published 5 September 2017)

We analyze subway arrival times in the New York City subway system. We find regimes where the gaps between trains are well modeled by (unitarily invariant) random matrix statistics and Poisson statistics. The departure from random matrix statistics is captured by the value of the Coulomb potential along the subway route. This departure becomes more pronounced as trains make more stops.

DOI: 10.1103/PhysRevE.96.030101

Random matrix statistics are expected to occur in a wide variety of interacting particle systems (see Ref. [1] for a review) and (1 + 1)-dimensional transportation models [2–6] are an important class of such systems. In Ref. [6], following the classical experimental result of Refs. [7,8], the authors proposed a mechanism for random matrix statistics in bus systems. In this Rapid Communication, we examine whether or not the New York City subway system (MTA) is well modeled by these statistics.

The bus system in Cuernavaca, Mexico in the late 1990's has become a canonical example of a system that is well modeled by random matrix theory (RMT) [7–10]. This bus system has a built-in, yet naturally arising, mechanism to prevent buses from arriving in rapid succession. This mechanism arises due to mutual competition from the drivers. Without this interaction, and mutual competition, one should expect that bus arrivals would be Poissonian [11]. Unlike these settings, the MTA has a globally controlled mechanism to space trains and eliminate collisions [12], which would suggest the significance of long-range effects. Nevertheless, the system is largely run manually [12] and it is thus natural to expect that the dynamics of the system is locally governed by interparticle interactions.

A natural signature of random matrix statistics is whether or not the spacing between particles at a given site obeys a Wigner surmise-type law. Consider the times $T$ between successive train arrivals at a given station, and consider the normalized spacing $\tau = T/\langle T\rangle$. Assuming that the system is well modeled by RMT statistics, one expects the spacing to satisfy

$$\frac{\#\{s \in \tau : s \leqslant t\}}{\#\tau} \approx \int_0^t \rho(s)ds, \rho(s) = \frac{32}{\pi^2} s^2 e^{-\frac{4}{\pi}s^2}, \qquad (1)$$

where $\langle\cdot\rangle$ represents the sample mean, the function $\rho(s)$ is known as the ($\beta = 2$) *Wigner surmise* (WS) [13], and $\#S$ gives the cardinality of the set $S$. This is the approximation of Wigner for the asymptotic ($N \to \infty$) gap distribution for successive eigenvalues in the bulk of an $N \times N$ Gaussian unitary ensemble (GUE) matrix [14]. This is computed by considering the $2 \times 2$ case. This approximation of Wigner agrees surprisingly well with the true limiting distribution as $N \to \infty$ [15].

Another natural statistic to consider is the *number variance*. Fix a time $T_0$ and consider the time interval $[T_0,T]$ for $T_0 \leqslant T \leqslant T_1$. Let $n(T)$ be the number of trains that arrive in this time interval. Once one has made many statistically independent observations of $n(T)$, the number variance is computed by

$$N(t) = \langle (n(T) - \langle n(T)\rangle)^2\rangle, \quad T = T_1\langle n(T_1)\rangle^{-1}t. \qquad (2)$$

This normalization is made so that $\langle n(T)\rangle \approx t$. The asymptotic prediction from RMT is

$$N(t) \approx \frac{1}{\pi^2}(\ln 2\pi t + \gamma + 1), \qquad (3)$$

where $\gamma$ is the Euler constant [13].

Here, we observe that (1) and (3) hold on a subset of the MTA system. We also find Poisson statistics within the MTA (which are also found in Puebla, Mexico [9]). For example, the southbound No. 1 train in northern Manhattan is well modeled by RMT statistics but the northbound No. 6 train is well modeled by Poisson statistics in the middle of its route. We also show that the train gap statistics tend to deviate more from RMT statistics as more stops are made. To quantitatively determine Poisson statistics versus RMT statistics we make the following ansatz for the (normalized to mean one) gap density function for $u \in [0,1]$,

$$p(s;u) := \int_0^s \frac{\rho(x/(1-u))}{1-u}\frac{e^{(x-s)/u}}{u}dx.$$

This is the density for the convex combination of an independent exponential and a WS random variable. A similar ansatz was used in Ref. [16] for an analysis of car spacing statistics. Using the Kolmogorov-Smirnov (KS) statistic we choose $u$ to fit this distribution to the data. A small value of $u$, combined with a small KS value, indicates RMT statistics. A value of $u$ near unity, and a small KS value, indicate Poisson statistics. We note that this transition (from RMT to Poisson) is also seen within RMT as the bandwidth of a Hermitian random matrix shrinks [17].

*Data collection.* Our data are obtained from the MTA real-time data feeds [18] that allow the user to obtain real-time train arrival times for many stations in the MTA system. Thus, our analysis has an advantage over that in Ref. [7] because the statistics of every station in the data feed can be analyzed. The stations can then be classified into those close to RMT statistics, Poisson statistics, or neither. Using the latitude and longitude coordinates of each station, which the MTA also provides, we can estimate the arclength of the subway track and analyze spatial distances. This is a component in our Coulombic analysis below.

[*] aukosh@math.utoronto.ca

[†] ttrogdon@math.uci.edu

2470-0045/2017/96(3)/030101(5)

We analyze the arrival times for the No. 1 and No. 6 trains. These trains operate on separate lines. The No. 1 train runs both northbound and southbound between Manhattan and the Bronx on the west side. The No. 6 train runs both northbound and southbound, also between Manhattan and the Bronx except on the east side. The stations at which the No. 1 train stops are labeled with integers between 101 and 140 [19], increasing from north to south. The same is true of the stations for the No. 6 train with integers ranging between 601 and 640.

We chose the No. 1 and No. 6 trains for the following reason. The MTA data feed provides data only for the lines No. 1–No. 6 and the midtown shuttle line. The shuttle line only has two stops, so we ignore it. The No. 1, No. 2, and No. 3 trains service a similar corridor in Manhattan. The first train is the "local" line in Manhattan and the second two are "express" lines and extend to Brooklyn. Similarly, the No. 4, No. 5, and No. 6 trains service a similar corridor in Manhattan, with the No. 6 train being the "local" line, and No. 4 and No. 5 being the "express" lines, which also run through to Brooklyn. Thus we choose the No. 1 and No. 6 as they are both "local" trains and run, to some extent, parallel to each other. It would be interesting to perform a similar analysis for the express trains, though we do not pursue this direction here.

Our data set consists of No. 1 and No. 6 train arrival times in seconds at all stations obtained on 48 days (39 weekdays) during the summer and fall of 2016. As we imagine that the working day hours, including "rush hours," are most relevant to most subway users, we only consider arrivals that occur between 8:00 A.M. and 6:00 P.M. on weekdays. For each station we have approximately 3500 arrivals. The MTA system keeps a minimum spacing between trains. To account for this, we subtract 90 s from every train gap. This number could be treated as a fitting parameter, but we keep it fixed. This leads to a small number of negative gaps. Then if $T$ is the collection of observed gaps (in seconds), define $\tau = (T - 90)/\langle T - 90\rangle$ to be the normalized train gaps.

*The Kolmogorov-Smirnov test.* Define the KS statistic [20]

$$\mathrm{KS}(u,\tau) := \sup_{t\in\mathbb{R}} \left| \frac{\#\{s \in \tau : s \leqslant t\}}{\#\tau} - \int_0^t p(s;u)ds \right|.$$

For $u = 0$, the null hypothesis is that the normalized gaps are distributed according the WS, and for $u = 1$, the null hypothesis is that the gaps are exponentially distributed. The KS test supposes that the samples are independent. From our data we obtain successive gaps which contain repeated data from the same train and are clearly not independent. To approximate independence, we only retain every fifth gap (approximately 30 min between samples) and we perform the KS test with approximately 700 samples. We consider the significance levels $\alpha = 0.01, 0.05, 0.1$ (low, moderate, and high significance, respectively). It follows from Refs. [21,22] that the null hypothesis cannot be rejected if ($u = 0, 1$)

$$\sqrt{\#\tau}\,\mathrm{KS}(u,\tau) < 1.62 \quad \text{when} \quad \alpha = 0.01,$$
$$\sqrt{\#\tau}\,\mathrm{KS}(u,\tau) < 1.35 \quad \text{when} \quad \alpha = 0.05,$$
$$\sqrt{\#\tau}\,\mathrm{KS}(u,\tau) < 1.22 \quad \text{when} \quad \alpha = 0.10.$$

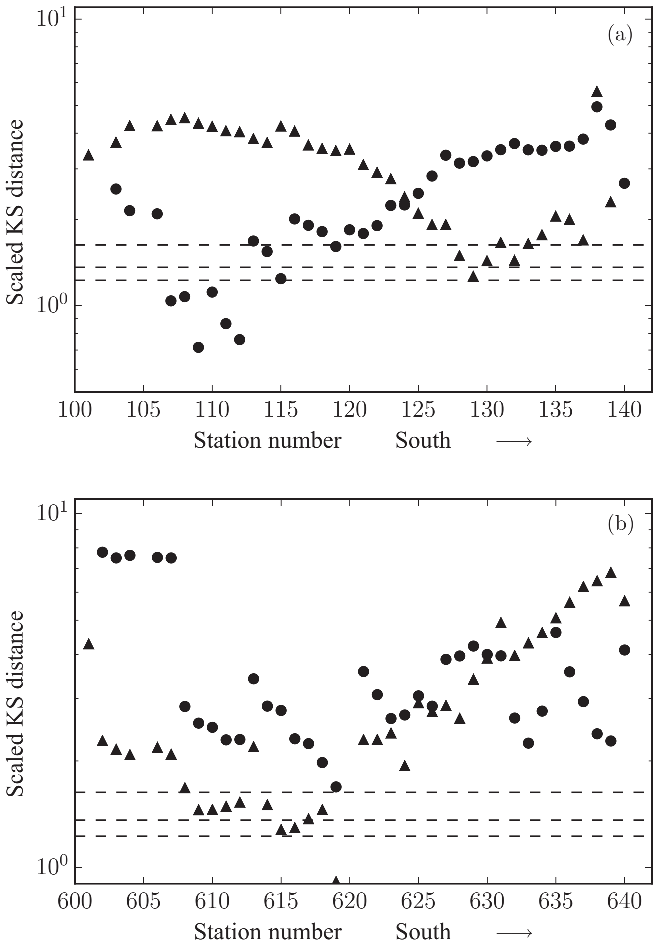

FIG. 1. The KS test for the No. 1 train (a) [$u = 0$] and the No. 6 train (b) [$u = 1$]. Circles and triangles represent southbound and northbound trains, respectively. The dashed lines in order of decreasing height represent the significance levels $\alpha = 0.01, 0.05, 0.1$. Stations that lie below a line pass the associated KS test.

In Fig. 1, we plot this scaled KS test statistic for every station on the northbound and southbound No. 1 and No. 6 trains. In particular, we find with high statistical significance ($\alpha = 0.10$) that six stations (107, 108, 109, 110, 111, 112) for the southbound No. 1 train pass the $u = 0$ KS test. If $\alpha$ is reduced, more stations pass the test. Similarly, for the northbound No. 6 train, one station passes the $u = 1$ KS test with high significance (619) and a total of three (615, 616, 619) stations pass the same test with moderate significance.

We emphasize that these tests are only suggestive of the underlying statistics. To illustrate this, consider the following experiment. Generate 2000 samples directly from the WS distribution. Then fit a (mean one) beta distribution [density proportional to $x^{\alpha-1}(1-x)^{\beta-1}$, then normalized to mean one] to the data by tuning the parameters $\alpha,\beta$ to minimize the KS statistic. Our experiments reveal that the data, with 2000 samples, are fit better by this tuned distribution than the WS distribution. So, one can never rule out such a beta distribution. Nonetheless, the KS statistic can be used to rule out Gamma distributions and the $\beta = 1,4$ Wigner surmises.

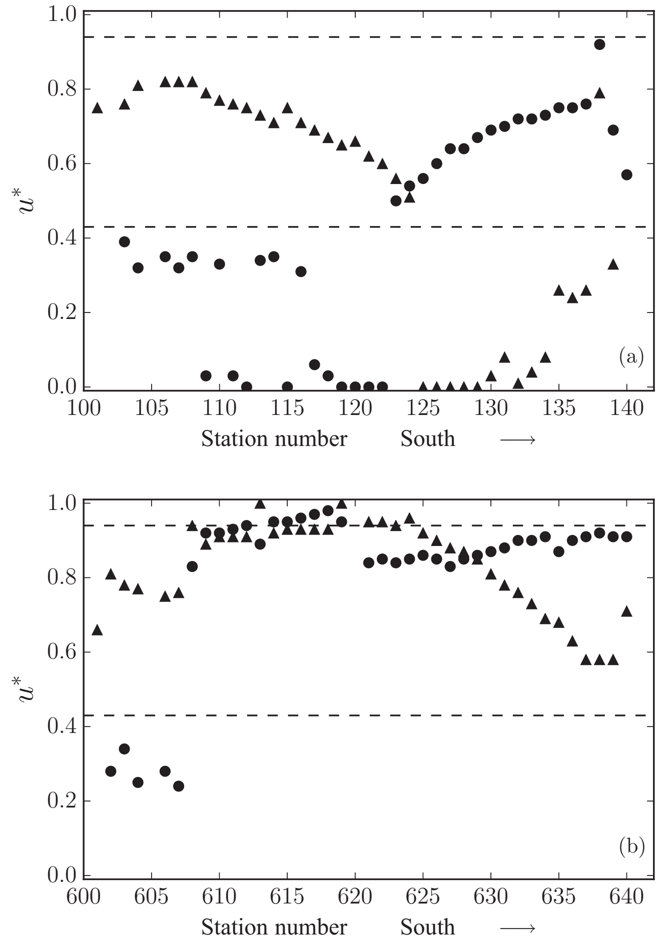

FIG. 2. The KS fit for the No. 1 train (a) [$u = 0$] and the No. 6 train (b) [$u = 1$]. Circles and triangles represent southbound and northbound trains, respectively. The dashed lines give the $u^* = 0.43$ and the $u^* = 0.94$ thresholds. Values of $u^*$ above 0.94 indicate Poisson statistics and values of $u^*$ below 0.43 indicate RMT-like statistics.

*A Kolmogorov-Smirnov fit.* The value of $u^*$ of $u$ that fits the data best is given by

$$u^* := \operatorname*{argmin}_{0\leqslant u\leqslant 1} \mathrm{KS}(u,\tau).$$

For every collection of normalized gaps $\tau$ this gives an optimal value $u^*$. Recalling that our sample sizes are approximately 700, we find that for $u < 0.43$ the KS test with moderate significance (comparing with $u = 0$) is passed. For $u > 0.94$ we find that the KS test with moderate significance is passed when comparing with $u = 1$. Stations with $u^* < 0.43$ are considered to exhibit RMT-like statistics and stations with $u^* > 0.94$ are considered to exhibit Poissonian statistics. The values of $u^*$ for each station and train are given in Fig. 2. These results should be compared with Fig. 1 to ensure significance. This presents further evidence that train gaps on the No. 1 train are RMT-like and those on the No. 6 train are Poissonian.

We choose station 112 and station 619 to examine in more detail. We display the normalized train gap histogram for both northbound and southbound trains at station 112 in Fig. 3. It is clear (and indeed highly statistically significant) that the southbound train gaps exhibit RMT-like statistics. But, in contrast, the northbound train appears to exhibit neither type of statistics. In Fig. 4, we display the normalized train gap histogram for northbound trains at station 619 which exhibits (with high statistical significance) Poissonian statistics.

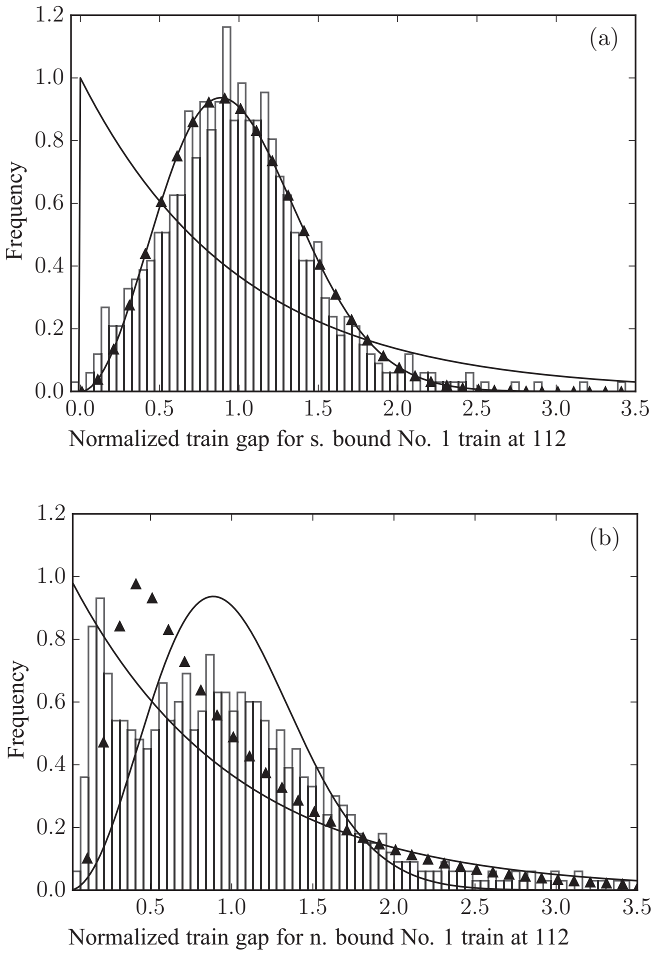

FIG. 3. The normalized train gap histograms for the northbound (bottom) and southbound (middle) No. 1 trains at station 112. The solid curves give the exponential and WS density. The triangles represent the best-fit density $p(s;u^*)$. The southbound train exhibits (highly significant) RMT statistics and our ansatz that determined $p(s;u^*)$ is not sufficient to capture the behavior of northbound trains.

*Number variance.* In order to compare with previous work, we also consider the number variance. As we are most interested in the RMT regime, we focus on the No. 1 train in stations that exhibit RMT statistics. To compute the number variance (2), we must obtain independent samples of the number of trains that arrive in a given time window. We record the arrivals of southbound No. 1 trains at stations 116 and 117 between 9:00 A.M. and 9:20 A.M. on weekdays. Our data limit us to 39 samples of $n(T)$. We plot the number variance against the theoretical prediction (3) in Fig. 5. While our agreement is not as good as that in Ref. [7], station 117 has good agreement for small values of $t$.

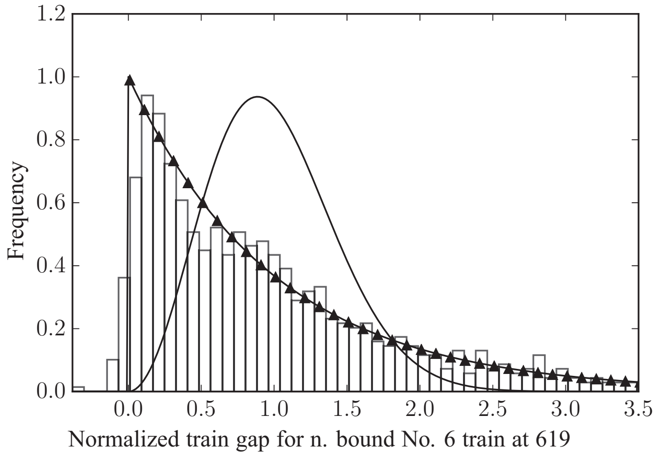

FIG. 4. The normalized train gap histograms for the northbound No. 6 trains at station 619. The solid curves give the exponential and WS density. The triangles represent the best-fit density $p(s;u^*)$. This station exhibits (highly significant) Poissonian statistics.

*The Coulomb potential.* The stationary distribution for an appropriately scaled ($\beta = 2$) Dyson Brownian motion is the distribution on the eigenvalues $\lambda_1 < \lambda_2 < \cdots < \lambda_N$ of a GUE matrix [23]. The Hamiltonian $H(\lambda) := \frac{1}{2}\sum_k \lambda_k^2 - \frac{1}{N}\sum_{j<k} \ln|\lambda_k - \lambda_j|$ is approximately conserved by the Dyson Brownian motion dynamics—the particle system $\lambda$ fluctuates near the minimum of this functional. The first term is referred to as the *confining potential*. Given the comprehensive information our data set gives us about the MTA system, we can plot many train trajectories simultaneously. Each train is represented by a function $\lambda_j(\ell)$ of the distance $\ell$ the train has traveled down the track. The value of $\lambda_j(\ell)$ is the time at which the train is a distance $\ell$ from its starting location. This is feasible using the latitude and longitude coordinates provided by the MTA for each station. The functional $H$, in a local sense, favors points that are regularly spaced. The minimizer of the functional is called the equilibrium measure and it is well studied in the literature [24]. It is natural to evaluate the functional to detect regularly spaced trains and a departure from random matrix statistics.

Each weekday, we monitor ten successive southbound No. 1 trains $\lambda_j(\ell)$, $j = 1,2,\ldots,10$, $0 \leqslant \ell \leqslant L$, starting with the first train ($j = 1$) that arrives at station 103 after 8:00 A.M. Each train is tracked until it reaches station 139. For each realization of these ten trains define

$$\mu_j(\ell) = \lambda_j(\ell) - 90j - \langle \boldsymbol{\lambda}(L) - \lambda(0)\rangle_j \frac{\ell}{L},$$

where the sample average $\langle\cdot\rangle_j$ is taken over $j$. This is used to estimate the "velocity" of the trains. Define the modified Coulomb potential

$$C[\boldsymbol{\mu}(\ell)] = -\sum_{j<k} \ln|\mu_k(\ell) - \mu_j(\ell)|. \tag{4}$$

Here, we drop the confining potential. We assume we are viewing the particle system on a microscopic scale and this potential is effectively constant. In Fig. 6 we plot the trajectories of $\mu_j(\ell)$ as a function of $\ell$ to demonstrate that the trains undergo nonintersecting motion. In Fig. 7 we plot the averaged Coulomb potential $\langle C[\boldsymbol{\mu}(\ell)]\rangle$, averaging over 29 weekdays [25]. The plot shows that the increase in the Coulomb potential is highly correlated with a larger scaled KS statistic. We can conjecture where the train statistics might be given by RMT based on the value of the Coulomb potential, presenting yet another connection to RMT.

It is worth pointing out in Fig. 7 that stations at a small distance fail the KS test but have a small Coulomb potential. This is largely from the fact that the fluctuations of the gaps are too concentrated about their means to agree with the WS. The trains start out at regularly spaced time intervals, nearly deterministic. As the trains progress down the track, larger fluctuations are introduced, giving rise to random matrix statistics while maintaining a small value of the averaged

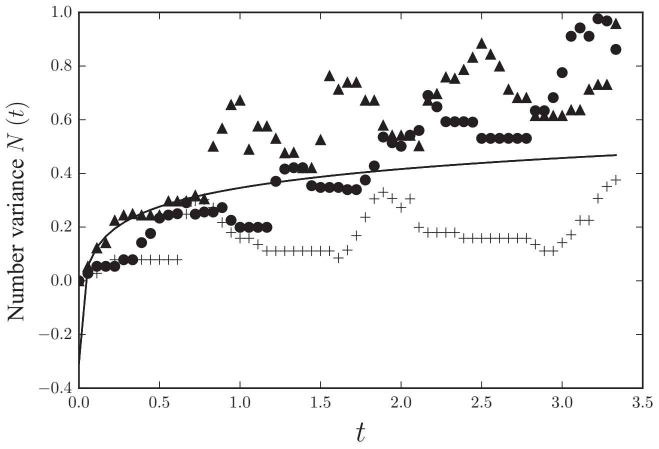

FIG. 5. The empirical number variance for southbound No. 1 trains at stations 116 (dots), 117 (triangles), and 112 (crosses) plotted against the theoretical curve (3). Agreement appears particularly good for station 117 for small values of $t$. Agreement is not as good for station 112.

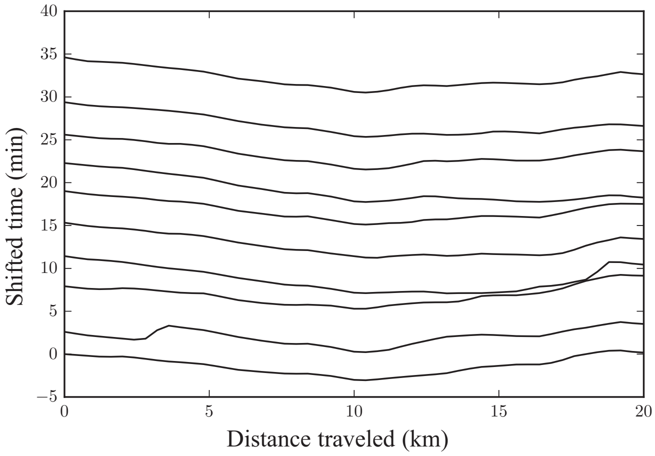

FIG. 6. Trajectories of the shifted southbound No. 1 trains $\mu_j(\ell)$, $j = 1,2,\ldots,10$. The horizontal axis represents the distance the train has traveled (measured from stop 103). Theses shifted trajectories are qualitatively similar to that of nonintersecting Dyson Brownian motion, at least for short distances.

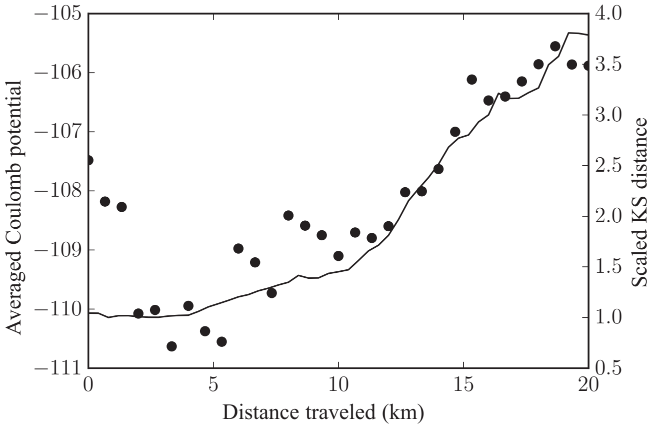

FIG. 7. The averaged Coulomb potential (4) for southbound No. 1 trains plotted as a function of distance from station 103. The scaled KS distance from WS is also plotted to show that when the Coulomb potential increases, so does the scaled KS statistic, indicating increased deviation from RMT statistics.

Coulomb potential. At some point, the external effects on the subway (such as passengers holding doors open) cause the system to depart from random matrix statistics, increasing the value of the averaged Coulomb potential. Thus the Coulomb potential, specifically its increase in value, is a mechanism for detecting less regularly spaced trains.

*Directions for further study.* The foregoing analysis focused on the behavior of "local" routes, as opposed to "express" routes. It would be interesting to explore if the statistical behavior of the trains exhibits variability between these choices. Furthermore, given the limited size of this data set, it would be of real interest to perform a similar analysis after observing the train systems on much longer time scales, such as for a year or longer. And if RMT-like statistics are desirable, the mechanism for their deterioration in the southbound No. 1 train should be investigated. Additionally, the nonappearance of RMT-like statistics in the No. 6 train is curious.

*Conclusion.* In summary, we have provided significant statistical evidence that the train gaps in the NYC MTA system exhibit RMT-like statistics. In addition, regimes exists where train arrivals are Poissonian. In this sense the MTA is a concrete physical system that exhibits both RMT and Poisson statistics. We have also used detailed spatial information to gain increased insight into the train correlations, treating their trajectories as those of a particle system. While we make no conjectures about the physical mechanisms behind the transition from RMT-like statistics to Poissonian statistics, RMT-like statistics do appear to be destroyed as the train makes more and more stops. But if one takes RMT-like statistics for train arrivals to be a hallmark of efficiency, as could be argued from the Cuernavaca, Mexico case study, this type of analysis may prove fruitful as a guide to understand and improve the performance of a subway system. The main conclusion of this Rapid Communication is that the "noise" of the subway system coming from train traffic and passengers can deteriorate purportedly beneficial statistical properties of the system. It is important to ask if the introduction of global computer control to the MTA will alleviate this issue.

*Acknowledgments.* The authors thank P. Deift for useful suggestions. This work was supported in part by NSF-DMS-1303018 (T.T.) and NSF-OISE-1604232 (A.J.). The authors also would like to acknowledge the support of the Courant Institute at New York University during the data collection portion of this project.

[1] T. Kriecherbauer and J. Krug, J. Phys. A: Math. Theor. **43**, 403001 (2010).
[2] F. Spitzer, Adv. Math. (NY) **5**, 246 (1970).
[3] L. Bertini and G. Giacomin, Commun. Math. Phys. **183**, 571 (1997).
[4] C. A. Tracy and H. Widom, Commun. Math. Phys. **290**, 129 (2009).
[5] B. Derrida, S. A. Janowsky, J. L. Lebowitz, and E. R. Speer, J. Stat. Phys. **73**, 813 (1993).
[6] J. Baik, A. Borodin, P. Deift, and T. Suidan, J. Phys. A: Math. Gen. **39**, 8965 (2006).
[7] M. Krbálek and P. Seba, J. Phys. A: Math. Gen. **33**, L229 (2000).
[8] M. Krbálek, P. Šeba, and P. Wagner, Phys. Rev. E **64**, 066119 (2001).
[9] M. Krbálek and P. Seba, J. Phys. A: Math. Gen. **36**, L7 (2003).
[10] M. Krbálek, J. Phys. A: Math. Theor. **41**, 205004 (2008).
[11] O. J. O'Loan, M. R. Evans, and M. E. Cates, Phys. Rev. E **58**, 1404 (1998).
[12] P. Dougherty, *Tracks of the New York City Subway* (Dougherty, New York, 2016).
[13] M. L. Mehta, *Random Matrices* (Academic, New York, 2004), p. 688.
[14] A GUE matrix is a Hermitian matrix with independent and identically distributed standard complex Gaussian entries, up to the symmetry condition.
[15] A numerical calculation using Fredholm determinants reveals that the KS distance is less than $5 \times 10^{-3}$.
[16] A. Y. Abul-Magd, Phys. Rev. E **76**, 057101 (2007).
[17] T. Shcherbina, Commun. Math. Phys. **328**, 45 (2014).
[18] MTA Real-Time Data Feeds (2016), http://datamine.mta.info/.
[19] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevE.96.030101 for a table to convert from station number to station name.
[20] $\#\tau$ is the cardinality of the set $\tau$.
[21] A. Kolmogorov, G. dell'Istituto Ital. degli Attuari **4**, 83 (1933).
[22] N. Smirnov, Ann. Math. Stat. **19**, 279 (1948).
[23] F. J. Dyson, J. Math. Phys. **3**, 1191 (1962).
[24] E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields* (Springer, New York, 1997), p. 509.
[25] Ten days were rejected because at least one of the chosen trains did not complete its route.