Les Cahiers du GERAD ISSN: 0711–2440

Estimating OD matrices from social connectivity: A per-origin probabilistic attractiveness model

R. Khalvandi, B. Sansò

G–2026–04

January 2026

La publication de ces rapports de recherche est rendue possible grâce au soutien de HEC Montréal, Polytechnique Montréal, Université McGill, Université du Québec à Montréal, ainsi que du Fonds de recherche du Québec – Nature et technologies.

Dépôt légal – Bibliothèque et Archives nationales du Québec, 2026
– Bibliothèque et Archives Canada, 2026

The publication of these research reports is made possible thanks to the support of HEC Montréal, Polytechnique Montréal, McGill University, Université du Québec à Montréal, as well as the Fonds de recherche du Québec – Nature et technologies.

Legal deposit – Bibliothèque et Archives nationales du Québec, 2026
– Library and Archives Canada, 2026

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Estimating OD matrices from social connectivity: A per-origin probabilistic attractiveness model

Reza Khalvandi
Brunilde Sansò

Département de génie électrique, Polytechnique Montréal & GERAD, Montréal, (Qc), Canada, H3T 1J4

reza.khalvandi-ilezoole@polymtl.ca
brunilde.sanso@polymtl.ca

January 2026
Les Cahiers du GERAD
G–2026–04

Abstract : Origin–destination (OD) matrices are essential for forecasting and capacity planning in transportation and communication networks, yet they are not directly observable and must be inferred from limited measurements. Classical gravity and power-law models capture broad distance effects but lack theoretical grounding and enough practical precision, while data-driven approaches require large training sets and are brittle under architectural change. We introduce a probabilistic framework that derives OD flows from individual interactions. The model separates end-to-end traffic into two observable components: how often individuals in two locations interact (their attractiveness) and the average traffic carried by each interaction. This decomposition provides a physically interpretable, data-fusable structure that remains valid regardless of network architecture. To evaluate the model, we focus on the empirically dominant case in which interaction probabilities decay with distance according to a power law. From this, we derive an origin-specific attractiveness measure in which the effective distance is scaled by the origin's population density and calibrated independently for each origin. Applied to large-scale county–pair social-interaction data, the resulting origin-specific fits substantially outperform classical, globally parameterized power-law models and recover exponents consistent with independent empirical estimates. These findings reinforce the core implications of the model: OD flows scale linearly with population, and the dominant source of heterogeneity stems from origin-specific interaction behavior rather than nonlinear population effects. This provides a practical pathway for robust, high-precision, end-to-end OD estimation, especially given the widespread availability of large-scale statistical records.

Keywords : Origin–destination (OD) matrices; gravity models; attractiveness; social connectedness; probabilistic OD modeling

1 Introduction

An origin-destination (OD) matrix quantifies the flow demand between origins and destinations in a network. In digital communication, this represents the volume of data traffic over a period, while in transportation, it represents the number of trips over specific intervals (e.g., hourly or daily). OD matrices are essential tools in network modeling, capacity planning, and forecasting across transportation [28] and communication networks [16, 25]. However, OD matrices are not directly observable; only aggregate link counts (total traffic on each link or road) are typically measured. Thus, estimating OD matrices from limited observations is inherently an under-determined inverse problem, as many matrices can produce identical observed link loads. To address this challenge, classical structural methods (gravity [12], power-law [40], entropy-based [36]) and data-driven methods [16, 28, 34, 38, 41] have been developed. Data-driven (AI-based) methods can achieve high precision, but they require large training data bulks, limiting their utility in network provisioning contexts where data are sparse or unavailable. Moreover, exclusive reliance on currently observed network data is increasingly problematic amid potential architectural shifts: in transportation (e.g., drones and other urban air mobility) and in digital communications (e.g., satellite backbones and heterogeneous access technologies). Under such evolving conditions, usage forecasts and origin–destination (OD) matrices cannot be reliably extrapolated from historical measurements alone. High-precision planning, therefore, requires end-to-end traffic measurements or models that remain valid under architectural change. In contrast, structural models exploit structural properties of OD flows, particularly their heavy-tailed distributions, often modeled using power-law decay [11]. Structural methods share a common principle: flow between two locations depends proportionally on their respective populations and inversely on a cost function related to distance or hops. Despite their simplicity, these models currently lack a strong theoretical foundation. Parameters are empirically determined without clear interpretation, creating ambiguity regarding their exact relationship to real-world conditions such as geography, topography, or social interactions, and consequently hampers efforts to improve their precision to a reliable level.

Addressing these gaps requires an OD framework that is grounded in the actual mechanisms through which flows arise. We develop such a framework by starting at the level of individual interactions. Each individual allocates a limited interaction budget across different applications, selects counterparts according to a probability distribution, and generates a measurable payload per application. Aggregating these microscopic interaction events across the population yields a macroscopic OD flow model in which traffic between two locations is the product of two conceptually distinct and observable quantities: (i) the number of interactions between the two locations, and (ii) the average traffic volume carried by each interaction. This decomposition replaces heuristic assumptions with physically interpretable parameters and connects OD estimation directly to digital datasets that measure social or communication interactions at scale. To evaluate how well this model aligns with real behavior, we focus on the empirically dominant case in which interaction probabilities decay with distance following a power-law. This decay pattern, repeatedly observed in large-scale datasets of social and digital interactions [4, 6, 7, 19–24, 31], allows a clean analytical derivation of origin-specific attractiveness. A key implication of this derivation is that attractiveness depends not only on distance but also on origin-specific characteristics—particularly population density and the distribution of interaction intensity—which together shape the mass and spatial intensity of interactions across both nearby and distant destinations. This origin-specific structure explains why classical gravity and globally parameterized power-law models systematically miss important heterogeneity in real OD flows. We validate these theoretical predictions using a large-scale county–pair dataset of social connectedness [5], which yields substantially stronger correlations between the model and observed interaction data compared with classical approaches. Our contributions are summarized as follows:

- We derive OD demand directly from individual interactions, separating total flow into the number of cross-location contacts and the average traffic per contact. This yields a simple and interpretable structure that remains valid under architectural change, and reveals that the apparent complexities in classical gravity and power-law models arise not from nonlinear population

effects but from origin-specific behavioral and socioeconomic factors. By identifying where heterogeneity truly resides (at the origin) and where structure is universal (linearity in population and probability normalization), the framework provides a principled basis for more accurate and extensible OD modeling.

- We use the dominant power-law decay in human interactions to derive an origin-specific attractiveness measure. It calibrates each origin separately by transforming actual distance into an effective distance through normalization with origin population density, with attractiveness intensity decaying as a function of effective distance from the origin. Applied to large-scale county–pair data [5], the results show that the proposed origin-specific formulation substantially improves correlation with observed interaction patterns compared to classical power-law models, particularly when each origin is calibrated independently. This demonstrates that the core implications of the probabilistic model—linearity with population, origin-specific normalization, and per-origin behavioral weighting—are indeed reflected in real-world data.

The rest of this paper is structured as follows: Section 2 reviews background and foundational concepts. Section 3 formally derives the probabilistic OD model. Section 4 formulates attractiveness for the case of a power-law distribution. Section 5 validates the theoretical model with empirical datasets. Finally, Section 6 summarizes findings and suggests avenues for future research.

2 Background

Mathematically, an OD matrix quantifies traffic volumes from origins to destinations, where each entry T_{ij} represents the flow from origin i to destination j. Unlike data-driven methods [16, 28, 34, 38, 41] that infer T_{ij} indirectly from aggregate link measurements, structural methods such as gravity and power-law models estimate T_{ij} directly based on theoretical assumptions and limited empirical data. The gravity model, inspired by Newton's law of gravitation, assumes the flow between two locations is proportional to their respective attractiveness (analogous to mass) and inversely proportional to a cost function, typically distance or travel impedance. Formally, the general gravity model is expressed as:

$$T_{ij} = K\, O_i^{\beta_i}\, D_j^{\beta_j}\, f(d_{ij}),$$

where O_i and D_j represent the attractiveness measures of origin i and destination j, respectively, d_{ij} represents the separation or effective distance between the pair, and K, β_i, and β_j are empirical model parameters. Often, for simplicity, $\beta_i = \beta_j = 1$, and the normalization constant K ensures the OD matrix aligns with the total observed flow. Common forms for the impedance function $f(\cdot)$ include exponential decay, $f(d_{ij}) = e^{-\alpha d_{ij}}$, and power-law decay, $f(d_{ij}) = d_{ij}^{-\alpha}$, with the parameter α derived from empirical data. Gravity models are attractive for their simplicity, minimal data requirements, and reasonable predictive accuracy. However, they inherently assume a smooth, monotonically decreasing relationship between distance and flow, which may oversimplify real-world complexities and variability. Power-law models explicitly recognize heavy-tailed distributions in flow patterns. Such models assume either that the distance impedance $f(\cdot)$ follows a power-law form or, more broadly, that the flow ranked by size follows a power-law distribution. For example, the k-th largest flow might follow a proportional relationship to $k^{-\alpha}$ for some exponent α. While power-law models capture the highly skewed nature of OD flows effectively, they typically require additional calibration or integration with other modeling techniques to accurately represent real-world scenarios. Devlin et al. [11], for instance, incorporated a preferential attachment mechanism into their model to better match observed traffic distributions. Despite their simplicity, both gravity and power-law models suffer from significant theoretical limitations. Key parameters like the normalization constant K and exponents β_i and β_j lack rigorous theoretical foundations. For instance, the gravity model does not provide clear justification as to why the constant K should remain uniform across all OD pairs, nor why exponents β_i and β_j might vary independently—an issue absent in Newtonian gravity analogues. Furthermore, the definition and measurement of effective distance or cost functions remain ambiguous, lacking a well-established

theoretical grounding. These deficiencies create an analytical impasse: heuristic formulations without a solid physical interpretation offer limited insight into how model accuracy or generality can be systematically improved for reliable use in network planning. In particular, the so-called structural models—including both classical gravity and its power-law variants—are not derived from the intrinsic mechanisms governing human or societal attractiveness and interaction. As a result, their parameters lack behavioral or physical meaning, making it difficult to refine or extend these models in a principled way. Without a formulation anchored in the fundamental nature of interaction between populations, any improvement remains empirical and ad-hoc rather than theoretically justified.

On the other hand, data-driven OD estimation methods often rely on large volumes of historical traffic data measured on a given network topology, which makes them notoriously *topology-dependent*. These models typically infer origin-destination demands from link-level measurements (e.g. counts on network links), implicitly assuming the current routing and network structure remain fixed [37, 39]. As a result, a model trained or calibrated on one network environment can fail to generalize when the network layout or technology changes [8]. In fact, simply feeding more historical data from the same topology does not overcome this limitation— the model still "learns" patterns tied to that specific network structure [37]. This sensitivity to historical topology and measurements is a critical weakness of data-driven approaches: their predictions and inferences degrade if the underlying graph of routes changes even moderately. The fundamental issue is that many more OD flows exist than direct observations, so inferring end-to-end traffic from aggregated link counts is an ill-posed problem without strong assumptions [29]. In other words, link-level data alone cannot always distinguish different origin-destination pairs, meaning a model trained on those link metrics is biased by the current topology and routing pattern.

These weaknesses become particularly problematic when designing new network topologies or technologies, where past link-level trends cannot be directly leveraged. For example, consider developing a global satellite communication network, which has a fundamentally different topology from the terrestrial Internet backbone, or a drone-based transportation system that relies on aerial routes instead of road networks. In such cases, using historical traffic measurements from existing Internet or road infrastructures can be misleading, as those measurements are inherently biased by the current terrestrial topology and aggregate many end-to-end flows along legacy routes [14]. They fail to reveal the true independent OD demands that would arise in the new infrastructure. To predict traffic under topology changes or in novel networks, one must incorporate end-to-end demand patterns and avoid topological bias. Advanced modeling frameworks instead attempt to estimate demand and routing jointly to capture how flows reroute or redistribute when the graph changes [37]. Such approaches implicitly recognize that pure data-driven models must either be topology-agnostic or retrained with new domain data to remain accurate in the face of network reconfigurations.

To address these limitations and develop a model that captures end-to-end traffic patterns while remaining topology-independent, we propose a framework grounded in the fundamental nature of human interaction. The model leverages statistical data on end-to-end human connectivity to predict OD flows. Specifically, we derive the entries of the origin–destination (OD) matrix by aggregating pairwise interactions between individuals in two entities (e.g., cities), where the total OD flow results from the sum of mutual attractiveness between origins and destinations multiplied by the average flow per end-to-end connection. Each parameter can be measured independently of the underlying network topology using the vast and diverse statistical datasets now available for different types of flows and applications. This formulation ensures that every model parameter has a clear physical or social interpretation, thereby enhancing its empirical measurability and theoretical transparency.

3 Social–interaction probabilistic OD modeling

To establish a solid theoretical foundation for OD flows, we propose a probabilistic model of individual-level interactions and derive macroscopic flows between entities (e.g., cities) as the aggregation of these microscopic interactions.

We fix a time period (e.g., one day) and let each individual k allocate a finite interaction *budget* across all potential counterparts and application types. For an application type $q \in \{1, \dots, q_{\max}\}$, individuals produce interaction events at an average rate γ^q (events per period), and each such event carries an average payload $\bar{T}^q$ (e.g., bytes in digital traffic).

We introduce non-negative weights w_k^q which represent the share of individual k's interaction budget devoted to application q. Since each person has limited resources (time, attention, money, etc.), the total budget allocated across all applications is bounded:

$$\sum_{q=1}^{q_{\max}} w_k^q \;\le\; W_0, \tag{1}$$

for some finite constant W_0 common to all individuals. The constraint in Equation (1) prevents any single individual from contributing an unbounded amount of flow to the aggregate OD traffic.

Conditional on generating a type-q event, individual k selects a counterpart l from the global population $\mathcal{L}$ according to probabilities P_{kl}^q, so that

$$\sum_{l \in \mathcal{L}} P_{kl}^q \;=\; 1. \tag{2}$$

The probability law constraint in Equation (2) induces a natural trade-off: for a given individual and application, increasing probability mass on nearby counterparts necessarily reduces probability mass on distant ones. In other words, local and remote interactions must share a common probability budget.

Let $M_j \subset \mathcal{L}$ denote the set of individuals belonging to city j. For a given origin individual k and application q, the probability that a type-q event is directed to some individual in city j is

$$\sum_{l \in M_j} P_{kl}^q.$$

Assuming that the payload per event is independent of the choice of counterpart, the expected type-q traffic generated by individual k towards city j over the period is

$$\mathbb{E}\big[T_{k\to j}^{(q)}\big] \;=\; \gamma^q \, \bar{T}^q \, w_k^q \sum_{l \in M_j} P_{kl}^q. \tag{3}$$

Aggregating over all applications yields the total expected traffic generated by individual k towards city j:

$$\mathbb{E}\big[T_{k\to j}\big] \;=\; \sum_{q=1}^{q_{\max}} \gamma^q \, \bar{T}^q \, w_k^q \sum_{l \in M_j} P_{kl}^q. \tag{4}$$

This individual-level description serves as the building block for our macroscopic origin–destination (OD) flow model. We then extend this framework to the macroscopic OD level to derive intercity flows. Specifically, instead of a single individual, we consider two cities with populations M_i and M_j. The total traffic flowing from City i to City j is obtained by summing the expected contributions of all individuals located in i:

$$T_{ij} = \sum_{k \in M_i} \mathbb{E}[T_{k\to j}] = \sum_{k \in M_i} \sum_{q=1}^{q_{\max}} \gamma^q \bar{T}^q \, w_k^q \sum_{l \in M_j} P_{kl}^q. \tag{5}$$

It is convenient to reorganize Equation (5) by grouping terms associated with the same application q. This yields the decomposition

$$T_{ij} = \sum_{q=1}^{q_{\max}} T_{ij}^q, \qquad T_{ij}^q = \gamma^q \bar{T}^q \sum_{k \in M_i} w_k^q \left(\sum_{l \in M_j} P_{kl}^q \right), \tag{6}$$

where T_{ij}^q denotes the OD traffic associated with application q. The inner summation $\sum_{l \in M_j} P_{kl}^q$ gives the probability that an event generated by individual k is directed to any individual in City j. Weighting this probability by w_k^q and aggregating over all individuals in i yields a natural measure of the attractiveness between the two cities:

$$\mathbf{E}(C_{ij}^q) \;=\; \sum_{k \in M_i} w_k^q \sum_{l \in M_j} P_{kl}^q. \tag{7}$$

We refer to $\mathbf{E}(C_{ij}^q)$ as the *type-q attractiveness* between i and j. It corresponds to the expected number (or expected rate) of type-q contacts between the two cities. With this definition, Equation (6) takes the compact form

$$T_{ij}^q \;=\; (\gamma^q \bar{T}^q)\, \mathbf{E}(C_{ij}^q), \tag{8}$$

which expresses the OD traffic of application q as the product of two statistically measurable and conceptually independent quantities:

- $\mathbf{E}(C_{ij}^q)$: the *expected number of contacts* between City i and City j;
- $\gamma^q \bar{T}^q$: the *average payload generated per period* for application q.

Aggregating again over all applications yields the macroscopic OD law

$$T_{ij} = \sum_{q=1}^{q_{\max}} (\gamma^q \bar{T}^q)\, \mathbf{E}(C_{ij}^q). \tag{9}$$

Equation (9) reveals a fundamental decomposition of OD traffic: flow volumes depend only on (i) the *attractiveness* between origins and destinations, and (ii) the per-contact payload of each application. Importantly, both components are observable in practice. Many modern datasets directly measure attractiveness, including counts of online friendships [3, 4, 6, 7, 21], intercity call and SMS volumes [19, 20, 24], and interaction frequencies in online games or other digital platforms [10, 22, 23, 31]. These datasets provide precisely the type of information captured by $\mathbf{E}(C_{ij}^q)$. Similarly, the average per-contact payload $\gamma^q \bar{T}^q$ can be estimated directly from a sample of connections (e.g., the mean number of passengers per trip, or the average data volume per digital interaction). Taken together, these observations imply that end-to-end traffic between any OD pair can be estimated entirely from interaction data and per-contact load statistics, without relying on assumptions about the underlying transmission network or routing structure.

The probabilistic formulation above establishes a direct bridge between measurable attractiveness $\mathbf{E}(C_{ij}^q)$ and OD traffic. As illustrated in Figure 1, once the expected number of application-q contacts between two cities is known, multiplying it by the per-contact payload $\gamma^q \bar{T}^q$ immediately yields the corresponding OD traffic T_{ij}^q. Thus, this decomposition provides a practical basis for predicting end-to-end OD flows even when the underlying transmission network or routing is unknown. In many cases, attractiveness data for one application q also serve as a proxy for other applications q' for which direct records are unavailable. This is because $\mathbf{E}(C_{ij}^q)$ captures the underlying social or interaction intensity between two regions, a structure that tends to be similar across applications. Hence, the factorization $T_{ij}^q = (\gamma^q \bar{T}^q)\, \mathbf{E}(C_{ij}^q)$ naturally supports extrapolation, as depicted in Figure 1, across applications—a capability not shared by traditional curve-fitting gravity or power-law models.

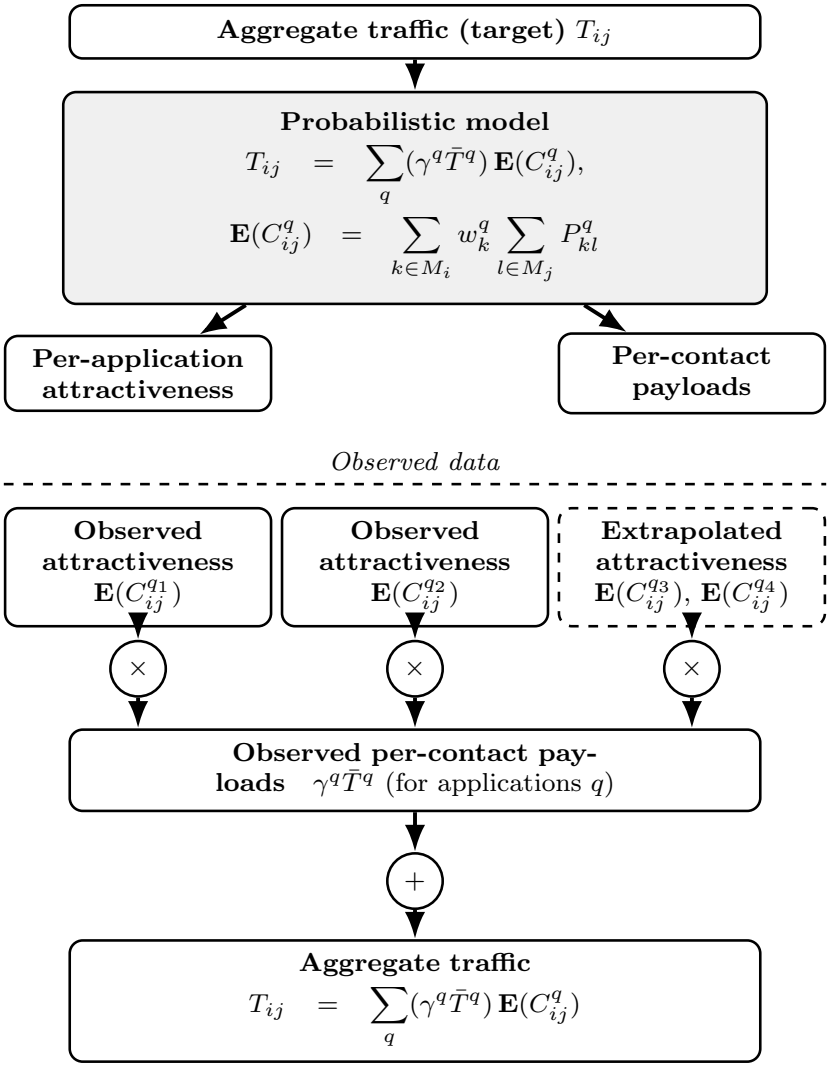

Figure 1: The probabilistic model separates OD flow into two measurable components: the expected attractiveness between origin and destination and the corresponding per-contact payload. Each observed or extrapolated attractiveness $\mathbf{E}(C^q_{ij})$ multiplies with its per-contact payload $(\gamma^q \bar{T}^q)$ (× nodes) to yield the traffic contribution of application q. These contributions are then summed (+) to produce the aggregate OD flow T_{ij}.

Analysis and the Implications of the model

To connect the proposed probabilistic framework with classical OD models, we adopt a standard behavioral assumption also underlying gravity and power-law formulations: interaction probabilities decay with distance. This assumption is widely used in the analysis of *wireless network capacity*—a distinct but mathematically related traffic modeling problem—where distance-based probabilistic interactions are central [2, 9, 13, 17, 18, 26, 35, 42]. Importantly, it is also strongly supported by extensive empirical evidence on human social interactions [4, 6, 7, 19–24, 31]. In most real-world OD settings, the distance between cities is large relative to their internal spatial extent. Under this geometric separation, individuals in City i perceive all individuals in City j as being approximately equidistant. As a result, the interaction probability between any individual pair $k \in M_i$ and $l \in M_j$ can be approximated by a common city-to-city interaction probability,

$$P^q_{kl} \approx P^q_{ij}, \qquad \forall\, k \in M_i,\ l \in M_j.$$

Substituting this approximation into the expression for OD traffic,

$$T_{ij} = \sum_{q=1}^{q_{\max}} \gamma^q \bar{T}^q \sum_{k \in M_i} w_k^q \left(\sum_{l \in M_j} P_{kl}^q \right)$$

$$\approx \sum_{q=1}^{q_{\max}} \gamma^q \bar{T}^q \, P_{ij}^q \, M_j \sum_{k \in M_i} w_k^q.$$

Introducing the origin-specific mean weight

$$\bar{w}_i^q \;=\; \frac{1}{M_i} \sum_{k \in M_i} w_k^q,$$

we obtain the simplified OD form

$$T_{ij} \;\approx\; \sum_{q=1}^{q_{\max}} (\gamma^q \bar{T}^q) \big(P_{ij}^q M_i M_j \bar{w}_i^q \big). \tag{10}$$

Equation (10) yields several important conclusions. In traditional gravity models, population exponents are treated as free parameters and are often fitted to values larger or smaller than one. In contrast, in our framework OD flows scale *linearly* with M_i and M_j, since population enters only through the explicit counting of individuals. Any departure from linearity must therefore arise from origin-specific application weights or interaction probabilities, $\bar{w}_i^q$ and P_{ij}^q, rather than from introducing nonlinear population exponents. In particular, the factor $\bar{w}_i^q$ captures the *interaction culture* of the origin, including the prevalence of specific applications and the socioeconomic structure of City i. Because these characteristics vary systematically across origins, $\bar{w}_i^q$ cannot be treated as a universal constant. In addition, the interaction probabilities P_{ij}^q are inherently origin-specific. For each origin i and interaction type q, they must satisfy the probability constraint

$$\sum_j P_{ij}^q = 1.$$

As a consequence, individuals (or origins) with highly concentrated local interactions must necessarily have fewer remote interactions, while individuals who interact more broadly must allocate less probability mass locally. This probability constraint produces a fundamental local–remote trade-off, conceptually illustrated in Figure 2, using a Gaussian distance decay distribution with standard deviation σ. For any radius r, the total probability mass can be decomposed as

$$\underbrace{\sum_{j:\, d(i,j) \le r} P_{ij}^q}_{\text{local mass}} + \underbrace{\sum_{j:\, d(i,j) > r} P_{ij}^q}_{\text{remote mass}} \;=\; 1.$$

As depicted in Figure 2, sharper kernels (smaller σ) allocate more probability locally and therefore reduce remote probabilities, while broader kernels do the opposite. This behavior is an unavoidable consequence of probability normalization and cannot be guaranteed by unconstrained gravity fits. It implies that local community structure fundamentally shapes long-range OD interactions.

In summary, the city–city approximation reveals that the key sources of heterogeneity in OD flows are origin-specific interaction patterns ($\bar{w}_i^q$) and origin-specific in distance-based probabilities (P_{ij}^q). These effects arise directly from the probabilistic foundations of the model and suggest that OD estimation methods should treat rows of the OD matrix on a per-origin basis rather than imposing global parameters across all entries. This perspective fundamentally changes how OD flows should be modeled. Classical gravity and power-law formulations typically introduce nonlinear population

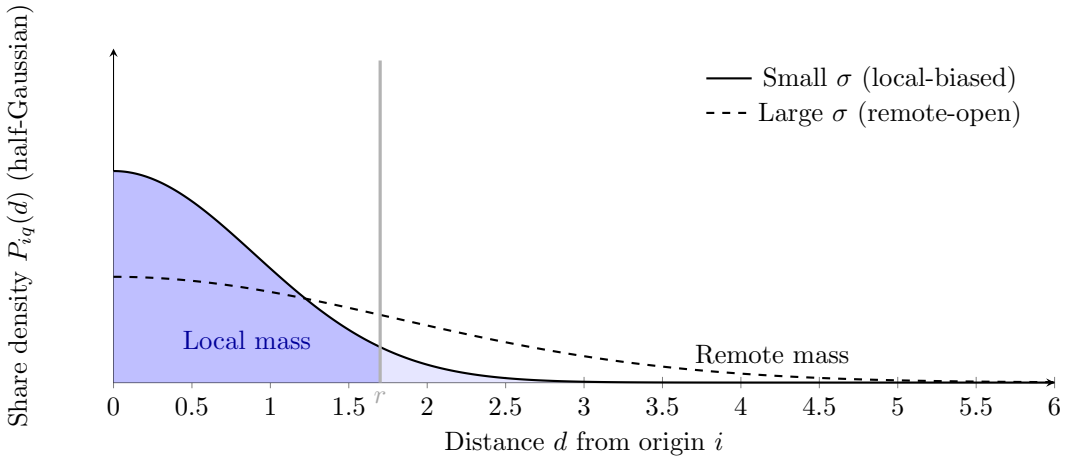

Figure 2: Origin-specific normalization with half-Gaussian share densities on $d \geq 0$. The smaller variance (solid) has a taller peak and concentrates probability locally; the larger variance (dashed) shifts mass to longer distances. For any radius r, the local and remote masses sum to 1.

exponents or universal scaling coefficients to fit empirical data. In contrast, the probabilistic model developed here shows that such complexity does not arise from population effects but from systematic per-origin differences in behavioral, technological, and socioeconomic factors encoded in $\bar{w}_i^q$ and in the origin-specific probabilities P_{ij}^q. Because these origin-specific terms stem directly from the underlying logic of individual human interactions, they provide a physically interpretable and extensible basis for model refinement. More detailed representations—incorporating heterogeneous application types, temporal variability, or demographic structure—can be added without altering the fundamental decomposition into expected contacts and per-contact payload. Thus, this probabilistic model not only enables end-to-end estimation of OD flows from digital traces but also offers a clear conceptual pathway for improving precision. Unlike purely empirical curve-fitting methods, it identifies where heterogeneity resides (at the origin) and where structure is universal (linear scaling in populations and normalization-constrained distance effects).

Finally, although comprehensive validation would ideally require joint access to interaction records, per-application payloads, and ground-truth OD flows, such datasets are rarely available together. Nevertheless, the model makes testable predictions that can be verified using more accessible social-interaction data. A central implication is the *linear relationship* between attractiveness and *OD traffic*, which allows attractiveness—much easier to measure than traffic—to serve as an effective proxy for flow. Empirical studies consistently show that intercity communication frequencies, online friendships, and other digital interactions exhibit stable, distance-decaying patterns that align closely with this structure [4, 6, 7, 19–24, 31]. Building on these observations, we develop a power-law attractiveness model in Section 4 and demonstrate that incorporating origin-specific features significantly improves fit relative to classical global power-law formulations. This provides meaningful empirical support for the probabilistic model: its key predictions—linear population scaling, per-origin heterogeneity, and normalized distance decay—yield substantially better alignment with real interaction patterns than traditional approaches.

4 Per-origin power-law attractiveness

Empirical studies consistently show that the frequency of social and digital interactions decays with distance according to a power law [4, 6, 7, 19–24, 31]. Consistent with this empirically established pattern, we derive in Section 4.1 an explicit per-origin attractiveness function under the assumption of power-law interaction probabilities. The resulting expression is a normalized attractiveness measure

that depends on distance, the power-law exponent, and a normalization factor enforcing that each individual's interaction probabilities sum to one. To express this normalization factor in terms of meaningful and observable quantities, we analyze it in detail in Section 4.2. This factor generally depends on several structural features, including city geometry, inter-city distances, the global population distribution, and the power-law exponent itself. We therefore examine a simplified yet generalizable scenario that isolates these dependencies and clarifies the role of normalization in shaping per-origin attractiveness.

4.1 Per-origin attractiveness over distance

Now, we compute the attractiveness around an origin for the special case of a power-law interaction probability between individuals. We consider the interaction probability of an individual ik with application q in origin i with any other person at distance d to follow

$$P_{ik}^{q}(d) = \frac{\lambda_{ik}^{q}}{d^{\alpha_q}}, \tag{11}$$

where λ_{ik}^{q} is a normalization constant ensuring that the total interaction probability for each individual sums to one, and α_q is the power-law exponent for the given application. As will be derived in Section 4.2 and supported empirically in Figures 5 and 6, under the assumption of a nearly uniform interaction opportunity over the area of origin i, the normalization constant λ_{ik}^{q} is approximately uniform across all individuals in city i. Therefore, we write

$$\lambda_{ik}^{q} \approx \lambda_{i}^{q}, \qquad \forall k \in M_i.$$

where $|M_i| = M_i$ is the population size of origin i. Moreover, as illustrated in Figure 3, we focus on the case in which the characteristic interaction distance d is much larger than the diameter of the focal city d_0. In this case, the spatial extent of origin i is negligible compared with d, and therefore all individuals in origin i effectively experience the same distance d to any location at that radius. To account for heterogeneity in the contribution of individuals to a given application, we associate to each individual $k \in M_i$ a non-negative weight w_k^q. Under this assumption, the weighted aggregate interaction probability of all individuals in origin i at distance d for application q becomes

$$\sum_{k\in M_i} w_k^q P_{ik}^q(d) \;\approx\; \sum_{k\in M_i} w_k^q \frac{\lambda_i^q}{d^{\alpha_q}} \;=\; \frac{\lambda_i^q}{d^{\alpha_q}} \sum_{k\in M_i} w_k^q. \tag{12}$$

In this definition, we introduce the average weight

$$\bar{w}_i^q \triangleq \frac{1}{M_i} \sum_{k\in M_i} w_k^q, \tag{13}$$

and rewrite the summation as

$$\sum_{k\in M_i} w_k^q P_{ik}^q(d) \;\approx\; \frac{M_i \bar{w}_i^q \lambda_i^q}{d^{\alpha_q}}. \tag{14}$$

Now, to compute the attractiveness of origin i, we consider a spatially continuous model as illustrated in Figure 3, where a focal city i is located at the origin with population M_i. Consider a thin annular ring centered at the origin with radius d and differential width δd. The area of this annulus is

$$\delta A = 2\pi d\, \delta d.$$

We assume that the population density at distance d in an annulus of width δd around the city is uniform and denote it by $\rho(d)$. Although this assumption is somewhat idealized, it underlies most power-law and gravity models. Under this assumption, the expected number of individuals in this differential ring is

$$\delta N = \rho(d) \cdot \delta A = 2\pi \rho(d)\, d\, \delta d.$$

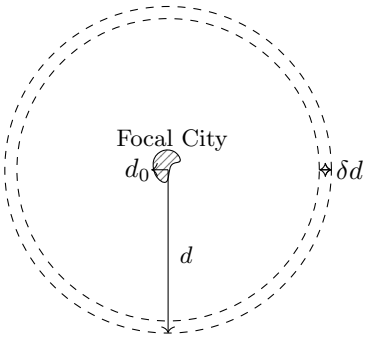

Figure 3: A focal city (origin) centered in a 2D plane. Individuals located in a thin annulus at distance d have the probability of interaction proportional to $d^{-\alpha_q}$.

Each individual in the ring interacts with origin i with probability given by the origin-level kernel in (14), so the expected attractiveness contributed by the ring, denoted $\delta C_i(d)$, for a city of population M_i is

$$\delta C_i^q(d) = \delta N \cdot \frac{M_i \bar{w}_i^q \lambda_i^q}{d^{\alpha_q}} = 2\pi\rho(d) M_i \bar{w}_i^q \lambda_i^q \, d^{1-\alpha_q} \, \delta d.$$

To account for spatial heterogeneity, we note that $\rho(d)$—the population density as a function of distance—can vary with location. To avoid the oversimplification of classical power-law and gravity models, which effectively assume a uniform environment, we normalize the term $\delta C_i(d)$ by $\rho(d)$. Consequently, the attractiveness is computed as

$$C_i^q(d) - C_0^q = \int \frac{\delta C_i(d)}{\rho(d)} = 2\pi M_i \bar{w}_i^q \lambda_i^q \int d^{1-\alpha_q} \, \delta d. \tag{15}$$

To compute $C_i^q(d)$ within distance d for $\alpha_q > 2$, we integrate from a reference distance with contact number C_0^q:

$$C_i^q(d) = M_i \bar{w}_i^q \, \frac{2\pi \, \lambda_i^q}{\alpha_q - 2} \, d^{2-\alpha_q} + C_0^q. \tag{16}$$

[18] has shown that, for any realistic power-law distribution, α_q should be strictly greater than two, which guarantees the convergence of the above integral. Taking the logarithm of both sides yields

$$\log\big(C_i^q(d) - C_0^q\big) = \log\left(\frac{M_i \bar{w}_i^q \, 2\pi \, \lambda_i^q}{\alpha_q - 2}\right) + (2 - \alpha_q) \, \log d \, . \tag{17}$$

Equation (17) shows that, in a log–log plot of per-origin attractiveness versus distance, the slope equals $2-\alpha_q$. Since the interaction probability itself is not directly observable from empirical data [18], most empirical studies instead report measures of attractiveness, contact distributions, or equivalent quantities. Consequently, when empirical studies fit a line to such data, the estimated exponent satisfies

$$\alpha_q = 2 - \text{slope}.$$

Hence, Equation (17) provides a crucial link between the theoretical interaction probability model and observational data for estimating the power-law exponent. Indeed, all measurements in studies such as [4, 6, 7, 19–24, 31] reflect contact distributions rather than direct interaction probabilities. Therefore, to avoid misleading interpretations, this distinction must be taken into account. For example, if the contact distribution decays as $1/d$, the corresponding interaction probability actually decays as $1/d^3$.

Equally important, even in an idealized setting where all origins share the same power-law exponent, heterogeneity persists due to the term $\log\left(\frac{M_i \, 2\pi \, \lambda_i^q}{\alpha_q - 2}\right)$, which induces origin-dependent offsets as $M_i \lambda_i^q \bar{w}_i^q$ varies. This effect is confirmed empirically in Section 5, where per-origin correlations at the best-fit α_q are substantially higher than those obtained from the full OD matrix. Consequently, unlike classical gravity or power-law models that impose a single global normalization, OD flows must be modeled at least on a per-origin basis, with each origin characterized by its own normalization constant.

4.2 Power-law probability normalization parameter

The probability normalization parameter is defined individually for each person such that the total probability of interaction across the entire population equals one. For example, if an individual S interacts, under application q, with any other individual v at distance d_v according to a power-law model $P_S^q(d_v) = \lambda_S^q / d_v^{\alpha}$, then

$$\sum_{v \neq S} P_S^s(d_v) = \sum_{v \neq S} \frac{\lambda_S^q}{d_v^{\alpha_q}} = 1.$$

To generalize this formulation spatially, we map each individual S to a physical location (x_0, y_0). There exists a one-to-one correspondence between individuals and their spatial coordinates, allowing us to express the normalization term as a spatially dependent function $\lambda^q(x_0, y_0)$ instead of λ_S^q.

Under this formulation, while the normalization condition is originally expressed as a discrete sum over all individuals, it can be approximated by a continuous integral over space. Specifically, we integrate the product of the interaction probability and the local population density $\rho(x, y)$ across the spatial domain centered at the individual's position (x_0, y_0):

$$\iint_{\text{domain}} P^q(x, y; x_0, y_0)\, \rho(x, y)\, \delta x\, \delta y = 1.$$

In this formulation, the population density $\rho(x, y)$ is assumed to be zero in unpopulated regions. To simplify the integration, we switch to polar coordinates centered at $(a\cos\theta_0,\, a\sin\theta_0)$. In these coordinates,

$$x = a\cos\theta_0 + r\cos\theta, \qquad y = a\sin\theta_0 + r\sin\theta\,.$$

For any point (r, θ), the interaction probability can be expressed as

$$P^q(r, \theta; a, \theta_0) = P^q(x, y; x_0, y_0) = \lambda^q(a, \theta_0)\, r^{-\alpha_q},$$

where $\lambda^q(a, \theta_0)$ is the normalization parameter at the reference point (a, θ_0). The differential area element is $\delta A = r\, \delta r\, \delta\theta$, and the population density at the corresponding position is

$$\rho(a\cos\theta_0 + r\cos\theta,\; a\sin\theta_0 + r\sin\theta).$$

Therefore, the normalization condition can be written explicitly as

$$\frac{1}{\lambda^q(a, \theta_0)} = \int_0^{2\pi} \int_{r_0}^{\infty} \rho(a\cos\theta_0 + r\cos\theta, a\sin\theta_0 + r\sin\theta)\, r^{1-\alpha_q}\, \delta r\, \delta\theta. \tag{18}$$

Conceptually, this integral sums up the contributions of all people in the plane, weighted by a distance-decay power-law factor. Points that are nearby contribute a term $\rho(a\cos\theta_0 + r\cos\theta, a\sin\theta_0 + r\sin\theta)\, r^{1-\alpha_q}$ which is relatively large (since r is small), whereas distant points contribute much less due to the $r^{-\alpha_q}$ decay. Furthermore, the integral will generally yield a larger value for points that are more centrally located in a city than for points near the city's edge, because a centrally located individual is surrounded by population in all directions. In contrast, an individual near the edge of a city has a portion of their surrounding circle falling outside the populated area (which contributes nothing).

Equation (18) presents a general formulation that can be applied to compute the normalization parameter of the power-law model for any spatial point. To gain deeper insight into how this parameter varies within a population center (e.g., a city) relative to the rest of the world, we now examine a simplified representation of a real-world scenario. This setup incorporates all relevant factors—including inter-city distances, global population distribution, and the power-law decay of interactions—within a tractable framework illustrated in Figure 4.

In this model, we consider a circular city of radius R and population M, surrounded by an empty annular gap extending from R to $R+d_g$, where d_g denotes the isolation distance separating the city from its surroundings. Beyond this gap lies another annular region representing the remainder of the global population, modeled as a uniform ring of thickness $d_{\max}$. We assume a constant population density ρ_1 inside the city, ρ_2 in the outer global region, and zero density in the intermediate gap. Consequently, the effective area contributing to the integral for $\lambda^q(a,\theta_0)$ comprises two distinct regions.

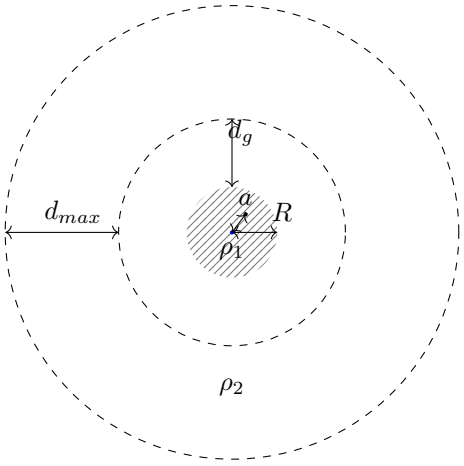

Figure 4: A central city of radius inscribed in an annular ring at distance d_g, which isolates it from the rest of the world.

As discussed in Theorem 2, it is sufficient to analyze a single representative directional arc—specifically, $\theta_0 = 0$, corresponding to the direction from the city boundary toward its center. The general approximation for this scenario is computed in Theorem 2 and given by Equation (19):

$$\begin{aligned}\frac{1}{\lambda^q(a,\theta_0=0)} = \frac{\pi}{6(\alpha_q-2)}\sum_{k=0}^{5}\Bigg[&\frac{2\rho_1}{r_0^{\alpha_q-2}}+\\ &\rho_2\Big(\sqrt{d_{max}^2+a^2\big(\frac{\cos^2(\frac{k\pi}{6})}{4}-1\big)}\pm\frac{a\cos(\frac{k\pi}{6})}{2}\Big)^{2-\alpha_q}\\ &-\rho_1\sqrt{R^2+a^2\big(\frac{\cos^2(\frac{k\pi}{6})}{4}-1\big)}\pm\frac{a\cos(\frac{k\pi}{6})}{2}\Big)^{2-\alpha_q}\\ &-\rho_2\Big(\sqrt{(R+d_g)^2+a^2\big(\frac{\cos^2(\frac{k\pi}{6})}{4}-1\big)}\pm\frac{a\cos(\frac{k\pi}{6})}{2}\Big)^{2-\alpha_q}\Bigg].\end{aligned} \tag{19}$$

As previously discussed, this integral serves as a continuous approximation of the discrete summation that accounts for the contribution of all individual interaction probabilities with respect to a specific individual. Therefore, the integration domain must exclude the area corresponding to the specific individual. To satisfy this condition, we set the lower limit of integration just outside the circle that contains exactly one individual. This condition implies $\rho_1\pi r_0^2 \geq 1$, and we consider the minimum value as the lower integral value as

$$r_0 = \sqrt{\frac{1}{\rho_1\pi}}. \tag{20}$$

We now perform a numerical analysis based on Equations (19) and (20) for the configuration illustrated in Figure 4, to examine how various factors influence the normalization parameter λ^q. First, we focus on the position change inside the city. In this regard, Figure 5 illustrates the variation of the value $\lambda^q(R-a,\theta=0)$ as a node shifts from the corner to the center of the city. In this setting,

we fix the densities ρ_1 and ρ_2 with $\rho_1 = \rho_2 = 42000$ per km^2, set the isolation gap to $d_g = 100$ km, and consider cities with populations ranging from 1,000 to 50 million. We perform this evaluation on several values of the power-law exponent α_q. Across all values of α_q, the expected behavior is observed: the value of λ^q is largest at the city corner due to the sparser surrounding population, but it quickly converges to $\lambda^q(0,0)$ as the node moves toward the city center. Interestingly, although we examine cities with vastly different populations (from 1,000 to 50 million), as long as the population density and α_q remain constant, the variation in λ^q is minimal. The curves nearly overlap and are difficult to distinguish. For instance, as highlighted for the case $\alpha_q = 3.4$, the ratio of λ^q between the most and least populous cities is approximately 99%. However, a small shift in the exponent—e.g., changing α_q by just 0.2—can cause a significant change in the λ^q value.

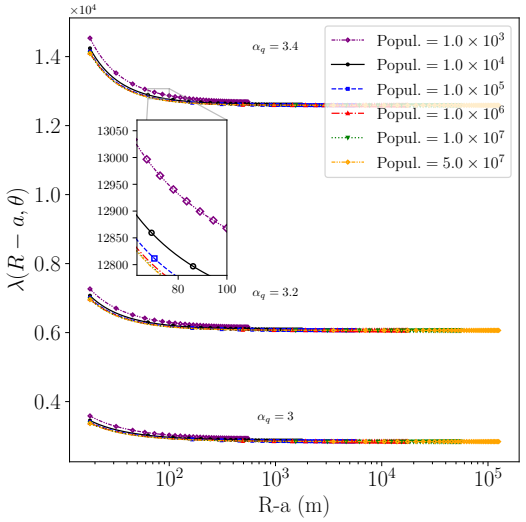

Figure 5: Effect of population size and power-law exponent on $\lambda^q(R-a,0)$ for $d_g = 100$ km and $\rho_1 = \rho_2 = 42{,}000$ persons/km^2, as a node moves from the city boundary toward its center, for cities with populations ranging from 1,000 to 50 million people.

Next, we perform another analysis using Equation (19), which covers the effect of the isolation distance d_g and population density on the value of λ^q. Figure 6 depicts the results of this numerical analysis, where the isolation distance d_g is varied from 1 km to 1000 km for a city with a fixed population of 100,000. As illustrated in Figure 6, the results indicate that the impact of isolation distance on λ^q is negligible, even when examined with a magnified view—no meaningful difference can be observed across the different values of d_g. In contrast, variations in population density exhibit a substantial influence on the value of λ^q, comparable to that of the power-law exponent α_q. These findings reinforce the conclusion that the dominant factors affecting λ^q are local parameters, specifically: (i) the population density ρ, and (ii) the distance decay parameter α_q. This holds true even though the full theoretical expression for λ^q may appear complex. Indeed, under the assumption of a sufficiently large population mass—implying that the contribution of the rest of the world to the interaction probability is negligible, which holds in most cases when the city diameter satisfies $R \gg r_0$—one can show, for an origin city i, from Equation (19), that

$$\lambda_i^q \propto (\alpha_q - 2)\, \rho_i^{-\frac{\alpha_q}{2}}, \tag{21}$$

given that $r_0 = \sqrt{\frac{1}{\rho_i \pi}}$ from Equation (20). This asymptotic relationship further emphasizes the dominant role of local population density and the power-law decay exponent in determining λ_i^q. Intuitively, under uniform population inter-city density, the normalization parameter follows the same power-law behavior as the underlying interaction kernel, scaling with the square root of the origin's population density.

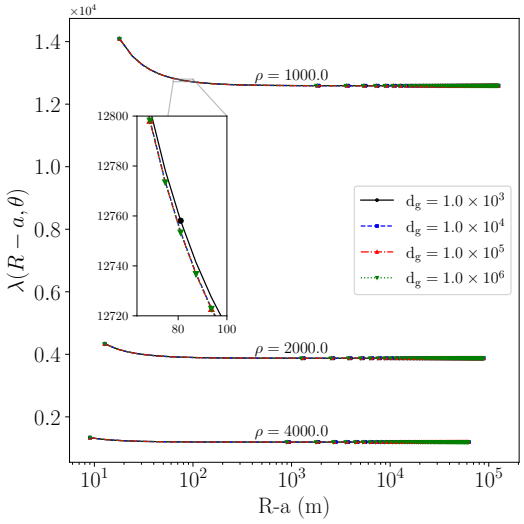

Figure 6: Effect of isolation distance d_g and population density on $\lambda^q(R-a,0)$. The isolation distance d_g is varied from 1 km to 1000 km for a city with a fixed population of 100,000, under different population densities ($\rho_1 = \rho_2$), as a node moves from the city boundary toward its center.

5 Data, validation and analysis

In this section, we examine the attractiveness power-law model introduced in Section 4 and validate it using a clean, high-coverage dataset: county-to-county Facebook friendship links across the entire United States, publicly available online [5]. The goal of this analysis is to test whether origin–destination attractiveness can be explained by the origin-specific structure embedded in OD flows and attractiveness. Because the probabilistic model presented in Section 3 predicts a linear relationship between flow and attractiveness, empirical alignment between the Social Connectedness Index data [5] and the theoretical attractiveness formulation in Section 4 would provide strong evidence that macroscopic attractiveness and flow patterns arise directly from microscopic individual interactions, consistent with the structure of our probabilistic model.

The Social Connectedness Index (SCI) dataset [5] provides a high-coverage, anonymized measure of social ties between nearly all pairs of U.S. counties. For an origin county i and destination county j, $\mathrm{SCI}_{ij} \geq 0$ captures the relative prevalence of social connections (e.g., Facebook friendships) linking residents of i to those of j, normalized by the total number of users in both counties. This normalization yields a scale-free index that is comparable across county pairs. Because SCI_{ij} is population-normalized, it is naturally aligned with the origin-normalized attractiveness $C_i(d_{ij})$ derived in Equation (19). Using the asymptotic form of λ_i^q from Equation (21), the attractiveness expression in Equation (19) implies

$$C_i^q(d_{ij}) \;\propto\; \bar{w}_i^q \frac{\rho_i\, \rho_i^{-\alpha_q/2}}{d_{ij}^{\alpha_q-2}} \;=\; \bar{w}_i^q \big(\sqrt{\rho_i}\, d_{ij}\big)^{2-\alpha_q}, \qquad \alpha_q > 0, \tag{22}$$

corresponding to a power-law decay in distance, normalized by the origin's population density. Because we do not impose a parametric model for $\bar{w}_i^q$, we focus on the minimal geographic component by defining the predictor

$$x_{ij}(\alpha_q) \;=\; \big(\sqrt{\rho_i}\, d_{ij}\big)^{2-\alpha_q}, \tag{23}$$

where d_{ij} is the great-circle distance between county centroids, ρ_i is the origin's population density, and $\alpha_q > 0$ is the decay exponent. The predictor $x_{ij}(\alpha_q)$ thus represents a density-normalized power-law distance term. We next evaluate how well this minimal geographic predictor aligns with empirical social connectivity patterns. To quantify agreement between theory and data, we compute the pooled Pearson

correlation between SCI_{ij} and a distance-based predictor over a grid $\alpha_q \in [2.01, 6.5]$. Specifically, we consider two predictors: the density-normalized power-law term $x_{ij}(\alpha_q)$ defined in Equation (23), and the unscaled distance term $d_{ij}^{2-\alpha_q}$. The corresponding correlations are

$$R_{\text{scaled}}(\alpha_q) = \operatorname{corr}\big(x_{ij}(\alpha_q),\, \mathrm{SCI}_{ij}\big), \tag{24}$$

$$R_{\text{raw}}(\alpha_q) = \operatorname{corr}\big(d_{ij}^{2-\alpha_q},\, \mathrm{SCI}_{ij}\big). \tag{25}$$

The analysis of the SCI data is conducted on a merged dataset comprising $N_c = 3{,}132$ counties (including county equivalents) and approximately $N_p \approx 10^7$ ordered county pairs after standard validity checks. County land area and centroid coordinates are obtained from the 2020 Census Gazetteer [32], while population counts are taken from the Census Population Estimates, Vintage 2024 [33]. Population density ρ_i is computed as population divided by land area. Self-pairs ($i = j$) are removed, and correlations are computed over valid county pairs only. Beyond these minimal joins and positivity constraints, we do not alter or reweight the inputs.

The retained sample satisfies basic harmonization conditions, including $i \neq j$, $d_{ij} > 10$ km, $\rho_i > 0$, and $\mathrm{SCI}_{ij} > 0$. The resulting correlations are reported in Figure 7. The scaled predictor exhibits a clear unimodal maximum,

$$\max_{\alpha_q} R_{\text{scaled}}(\alpha_q) \approx 0.596 \quad \text{at} \quad \alpha_q^\star \approx 3.51,$$

whereas the unscaled distance achieves a substantially lower peak correlation,

$$\max_{\alpha_q} R_{\text{raw}}(\alpha_q) \approx 0.365 \quad (\text{near } \alpha_q \approx 3.3).$$

This gap demonstrates that normalizing distance by origin density ρ_i effectively captures the dominant source of cross-origin heterogeneity (Section 4.2), aligning geographic separation with each origin's intrinsic interaction scale. Our predictor $x_{ij}(\alpha_q)$ incorporates this effect by replacing the raw distance d_{ij} with the density-scaled term $\sqrt{\rho_i}\, d_{ij}$. This formulation arises naturally from the the per-origin attractiveness analysis in Section 4. Although the raw distance term $d_{ij}^{2-\alpha_q}$ also exhibits a clear unimodal correlation peak consistent with power-law behavior, such heuristic models—lacking a probabilistic foundation—fail to capture the phenomenon with comparable precision.

Moreover, empirically, the $R_{\text{scaled}}(\alpha_q)$ behavior in Figure 7 is consistent with prior slope estimates. For the same dataset, [4] report a contact–distribution slope of -1.48. Using the mapping in [18], this corresponds to $\alpha_q \approx 3.48$, which is very close to our data–driven optimum $\alpha_q^\star = 3.51$.

To further illustrate the effect of origin-specific scaling, Figure 8 compares the distance dependence of the raw Social Connectedness Index (SCI) with that of a density-scaled SCI. The scaled SCI is defined as $\mathrm{SCI}_{ij}(\rho_i/\overline{\rho})^{(\alpha_q^\star - 2)/2}$, where ρ_i denotes the population density of the origin county and $\overline{\rho}$ is the mean U.S. population density. In both cases, median values are computed within logarithmically spaced distance bins and displayed on log–log axes. The comparison highlights how origin-density normalization modifies the observed distance-decay behavior. Under an ideal power-law relationship, the plot is expected to be linear with a slope of $\alpha_q^\star - 2$. In practice, heterogeneity induces curvature (and even peaks) in the unscaled case, whereas the density–scaled SCI aligns much more closely with a straight line and is well approximated by the theoretical slope. This improvement reflects the role of density normalization in absorbing per–origin diversity, particularly at shorter distances where destinations share similar environmental and social features with their origins. The *scaled* SCI clearly captures this effect. An important observation is that, in a power-law model, most of an origin's connections are concentrated at nearby destinations. Consequently, for estimating attractiveness or OD matrices, these short-range entries are the most critical and must be determined with high precision.

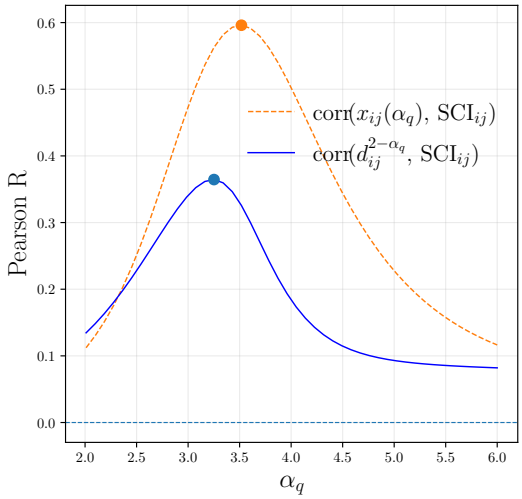

Figure 7: Pooled Pearson correlation $R(\alpha_q)$ over a grid of α_q for two predictors: $R_{\text{scaled}}(\alpha_q) = \mathrm{corr}\big(x_{ij}(\alpha_q), \mathrm{SCI}_{ij}\big)$ and $R_{\text{raw}}(\alpha_q) = \mathrm{corr}\big(d_{ij}^{2-\alpha_q}, \mathrm{SCI}_{ij}\big)$. The scaled predictor, which incorporates origin density, peaks at $\alpha_q^\star = 3.51$ with $R_{\max} = 0.60$, while the unscaled predictor exhibits a consistently lower correlation across all α_q.

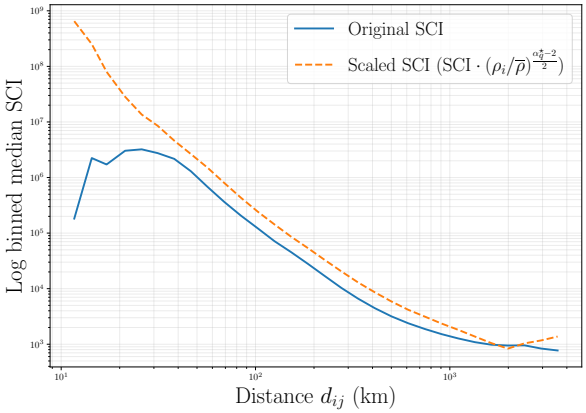

Figure 8: Median slope of the Social Connectedness Index (SCI) versus distance, shown on log–log axes for both the raw and scaled cases. The scaled SCI is computed as $\mathrm{SCI}_{ij} \times (\rho_i/\overline{\rho})^{(\alpha_q^\star-2)/2}$, where ρ_i is the population density of origin i and $\overline{\rho}$ is the mean U.S. population density.

Considering the observed dependence of attractiveness on origin characteristics, we compute within–origin correlations R_i using vectorized group statistics to reveal robust distance–decay behavior. Specifically, for each origin i, we evaluate

$$R_i \;=\; \mathrm{corr}_j\big(x_{ij}(\alpha_q^\star), \mathrm{SCI}_{ij}\big),$$

where R_i measures how closely the empirical social–connectedness patterns around each origin follow the theoretical predictor at the optimal exponent $\alpha_q^\star$. Figure 9 shows the distribution of these per–origin correlations between SCI_{ij} and the density–scaled predictor $x_{ij}(\alpha_q^\star)$. Figure 9 shows that the distribution of R_i is substantially higher than the pooled (all–pairs) correlation, with most counties exhibiting consistently high values (mean $R \approx 0.8$). This indicates that OD behavior becomes considerably more predictable once conditioned on the origin, underscoring the importance of origin-specific factors such as $\bar{w}_i^q$. This is fully consistent with the role of origin-level scaling in absorbing heterogeneity, as discussed in Section 4. Such per–origin treatment is absent in classical power-law and gravity models. Furthermore, although the density–scaled predictor $x_{ij}(\alpha_q)$ markedly improves estimation precision compared to heuristic power-law models, there remain important real-world factors influenc-

ing attractiveness that are difficult to capture theoretically. These factors, while partially calibrated through per–origin analysis, highlight the essential role of large-scale statistical data in bridging this gap and achieving higher practical accuracy.

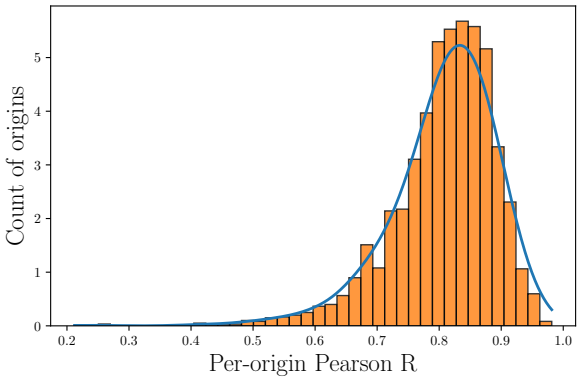

Figure 9: Distribution of per-origin correlations between SCI and $\left(\sqrt{\rho_i}\, d_{ij}\right)^{2-\alpha_q^\star}$. The mean R value is approximately 0.8.

In other words, although the density–scaled predictor $x_{ij}(\alpha_q)$—derived from the power-law interaction model in Section 4—demonstrates that the precision of classical formulations such as the power-law and gravity models can be enhanced, there also exist alternative or complementary approaches to address geographic heterogeneity that affects travel difficulty between cities and, consequently, their mutual attractiveness. For instance, some models calibrate effective distance using measures such as driving time or accessibility. Yet the number of effective factors in social interaction extends far beyond geography itself: [4] shows that societal boundaries shaped by cultural and political contexts must also be accounted for. Consequently, there will always be limitations in theory to capture all influential factors. A practical path to higher, truly applicable precision is therefore to leverage contemporary statistical evidence from diverse digital interaction sources, now widely available; such data naturally absorb the complex effects of multiple factors beyond idealized theory. Furthermore, social-interaction data are no longer limited to a few classic social networks. Today, many diverse sources extend beyond social media—for example, routing and navigation apps, ride-hailing platforms, and online delivery services—which can reveal end-to-end traffic flows in transportation. Combining these heterogeneous statistical data sources on social interaction can substantially improve the precision of end-to-end OD flow estimation. At the same time, a well-defined theoretical framework remains essential: it clarifies the structure of traffic patterns and provides principled guidelines for designing and interpreting measurements in social-interaction datasets.

6 Conclusion

We developed a probabilistic OD framework that derives end-to-end traffic flows between cities directly from individual-level interactions. In this framework, traffic between an origin and a destination is proportional to two quantities: how frequently individuals in those locations interact (their "attractiveness") and the average amount of traffic generated by each interaction. The model explains why OD flows scale linearly with origin and destination populations and shows that the true source of complexity lies in origin-specific behavior, rather than in arbitrary nonlinear population exponents. To test the framework, we focused on the empirically dominant case in which interaction probabilities decay with distance according to a power law and derived a corresponding per-origin attractiveness model. Applied to large-scale county–pair data, this formulation yields substantially higher correlations with observed interaction patterns than classical power-law models based on a single global parameter set. In other words, the core implications of the model—population linearity and per-origin calibration—are clearly reflected in real data and lead to markedly improved explanatory power. This framework thus

opens a new avenue for high-precision end-to-end OD flow prediction by leveraging large-scale digital traces of communication and transportation interactions, independent of the underlying communication network. At the same time, the framework remains open to further refinement. It highlights three fundamental ingredients shaping traffic: the probability that individuals in different locations interact, the intensity of those interactions, and the amount of traffic generated per interaction. Origin-specific interaction intensities, in particular, capture the "interaction culture" of each city and are themselves objects of interest for the social and behavioral sciences. Improved empirical characterization of these components can therefore advance OD prediction while preserving a transparent and interpretable modeling foundation.

Appendix A

Theorem 1. If we consider any point inside a circle of radius R, and divide the circle into 12 sectors, each of them covers $\pi/6$ with respect to that point, then using the geometry of the plane, the distances r_{1k} and r_{2k} can be defined as:

$$r_{1k}, r_{2k} = \sqrt{R^2 + a^2\left(\frac{\cos^2(\theta+\varphi_k)}{4} - 1\right)} \pm \frac{a\cos(\theta+\varphi_k)}{2},$$

Proof. Consider a point inside a circle with radius R, and divide the circle into 12 sectors, each covering $\pi/6$. First, consider the distances r_1 and r_2 in a pair of sectors. as shown in Figure A.1. For r_1 and r_2, we have:

$$(r_1 + a\cos(\varphi+\theta))^2 + a\sin(\varphi+\theta) = R^2\,, \quad \text{and}$$
$$(r_2 - a\cos(\varphi+\theta))^2 + a\sin(\varphi+\theta) = R^2$$

By summing these equations, we obtain,

$$r_1^2 + r_2^2 + 2a^2 + 2(r_1 - r_2)\cos(\varphi+\theta) = 2R^2\,. \tag{A.1}$$

Due to symmetry, we have:

$$r_2 - a\cos(\varphi+\theta) = r_1 + a\cos(\varphi+\theta)\,,$$

which implies

$$r_2 - r_1 = 2a\cos(\varphi+\theta)\,. \tag{A.2}$$

Solving the set of Equations (A.1) and (A.2)for r_{1k} and r_{2k} for any pair of sectors we get

$$r_{1k}, r_{2k} = \sqrt{R^2 + a^2\left(\frac{\cos^2(\theta+\varphi_k)}{4} - 1\right)} \pm \frac{a\cos(\theta+\varphi_k)}{2}, \tag{A.3}$$

where $\varphi_k \in \{0, \cdots, \frac{(k-1)\pi}{6}, \cdots, \frac{5\pi}{6}\}$. □

Theorem 2 ($\lambda^q(a,\theta)$ estimation)**.** Consider a circular city with radius R, internal node density ρ_1, surrounded by a region extending from a guard distance d_g up to a maximum distance d_{max} with density ρ_2, as shown in Figure 4. For a node located at a polar coordinate $(a, \theta_0 = 0)$ relative to the city center, the interaction normalization constant $\lambda^q(a, \theta_0 = 0)$ in the power-law case can be estimated by Equation (A.4).

Proof. We first divide the area around a node at position (a, θ_0) into 12 equal angular sectors, each with angle $\pi/6$. Due to symmetry, the integral of Equation (19) over these sectors with uniform internal density ρ_1 is expressed as:

$$\sum_{k=1}^{12} \frac{\pi}{6} \int_{r_0}^{r_k} \rho_1 r^{1-\alpha} \delta r.$$

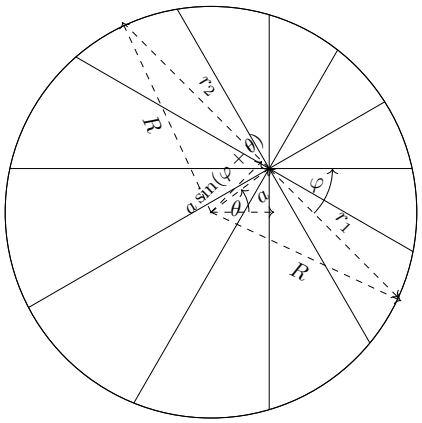

Figure A.1: Geometry for deriving the radial boundary distances r_{1k} and r_{2k} from an interior point at offset (a, θ) inside a circle of radius R. The circle is partitioned into 12 equal sectors of angle $\pi/6$. For each sector pair symmetric about the center, the distances to the circle boundary are measured along two opposite rays at angles $\theta + \varphi_k$ and $\theta + \varphi_k + \pi$, with $\varphi_k \in \{0, \pi/6, \ldots, 5\pi/6\}$.

Due to symmetry, it suffices to evaluate at $\theta_0 = 0$, giving radial r_{1k}, r_{2k} calculated from Equation (A.3) for $0 \leq k \leq 5$ as :

$$r_{1k}, r_{2k} = \sqrt{R^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}.$$

Evaluating the radial integral explicitly:

$$\int_{r_0}^{r_k} \rho_1 r^{1-\alpha} dr = \frac{\rho_1}{\alpha - 2}\left(\frac{1}{r_0^{\alpha-2}} - \frac{1}{r_k^{\alpha-2}}\right).$$

Therefore, the integral inside the city in Figure 4 becomes:

$$\frac{\rho_1 \pi}{6(\alpha-2)} \sum_{k=0}^{5} \frac{2}{r_0^{\alpha-2}} - \left(\sqrt{R^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha}.$$

For the outer region (from $R + d_g$ to d_{max}), using density ρ_2, the integral similarly gives:

$$\frac{\rho_2 \pi}{6(\alpha-2)} \sum_{k=0}^{5} \left(\left(\sqrt{d_{max}^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha} - \left(\sqrt{(R+d_g)^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha}\right).$$

Combining these inner and outer integrals yields the total inverse normalization constant $\frac{1}{\lambda^q(a,\theta)}$:

$$\begin{aligned}\frac{1}{\lambda^q(a, \theta_0 = 0)} = \frac{\pi}{6(\alpha-2)} \sum_{k=0}^{5} \Bigg[& \frac{2\rho_1}{r_0^{\alpha-2}} + \\ & \rho_2\left(\sqrt{d_{max}^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha} \\ & - \rho_1\left(\sqrt{R^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha} \\ & - \rho_2\left(\sqrt{(R+d_g)^2 + a^2\left(\frac{\cos^2(\frac{k\pi}{6})}{4} - 1\right)} \pm \frac{a\cos(\frac{k\pi}{6})}{2}\right)^{2-\alpha}\Bigg]. \quad \text{(A.4)}\end{aligned}$$

□

References

[1] Thomas J Allen. Communication and organizational structure. Technology Review, 79(1):38–49, 1977.

[2] Bita Azimdoost, Hamid R Sadjadpour, and JJ Garcia-Luna-Aceves. Capacity of wireless networks with social behavior. IEEE Transactions on Wireless Communications, 12(1):60–69, 2012.

[3] Lars Backstrom, Eric Sun, and Cameron Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In Proceedings of the 19th international conference on World wide web, pages 61–70, 2010.

[4] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness: Measurement, determinants, and effects. Journal of Economic Perspectives, 32(3):259–280, 2018.

[5] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. Social connectedness index (sci). Facebook Data for Good / Meta; county–county and international variants, 2021. Access via HDX or Data for Good portals.

[6] Michael Bailey, Patrick Farrell, Theresa Kuchler, and Johannes Stroebel. Social connectedness in urban areas. Journal of Urban Economics, 118:103264, 2020.

[7] Michael Bailey, Drew Johnston, Theresa Kuchler, Dominic Russel, Bogdan State, and Johannes Stroebel. The determinants of social connectedness in europe. In Social Informatics, pages 1–14, Cham, 2020. Springer International Publishing.

[8] Yumin Cao, J.W.C. van Lint, Panchamy Krishnakumari, and Michiel C.J. Bliemer. Data-driven origin–destination matrix estimation on large networks: A joint od–path choice formulation. Transportation Research Part C: Emerging Technologies, 168:104850, 2024.

[9] Zairan Cheng and Ying Liu. Performance study and optimization of 3d-manet:: A new analytical perspective based on zipf's law. Wireless Communications and Mobile Computing, 2022, 2022.

[10] Galina Daraganova, Pip Pattison, Johan Koskinen, Bill Mitchell, Anthea Bill, Martin Watts, and Scott Baum. Networks and geography: Modelling community network structures as the outcome of both spatial and network processes. Social networks, 34(1):6–17, 2012.

[11] Pat Devlin, Jeremy Kepner, Ashley Luo, and Erin Meger. Hybrid power-law models of network traffic. In Proc. IEEE Int. Parallel and Distributed Processing Symp. Workshops (IPDPSW), 2021.

[12] Vasile Dragu and Eugenia A. Roman. The origin–destination matrix development. In MATEC Web of Conferences, volume 290, page 06010, 2019.

[13] Luoyi Fu, Wentao Huang, Xiaoying Gan, Feng Yang, and Xinbing Wang. Capacity of wireless networks with social characteristics. IEEE Transactions on Wireless Communications, 15(2):1505–1516, 2015.

[14] Xinyu He, Yanfang Mo, Jiawei Huang, Lishuai Li, and S. Joe Qin. A distributed route network planning method with congestion pricing for drone delivery services in cities. Transportation Research Part C: Emerging Technologies, 160:104536, 2024.

[15] Yanqing Hu, Yougui Wang, Daqing Li, Shlomo Havlin, and Zengru Di. Possible origin of efficient navigation in small worlds. Physical Review Letters, 106(10):108701, 2011.

[16] Weiwei Jiang. Internet traffic matrix prediction with convolutional lstm neural network. Internet Technology Letters, 5(2):e322, 2022.

[17] Mohsen Karimzadeh Kiskani, Bita Azimdoost, and Hamid R. Sadjadpour. Effect of social groups on the capacity of wireless networks. IEEE Transactions on Wireless Communications, 15(1):3–13, 2016.

[18] R Khalvandi and B Sanso. The feasibility of ultra large-scale distributed networks in symmetrical network typologies. Les Cahiers du GERAD ISSN, 711:2440, 2024.

[19] Gautier Krings, Francesco Calabrese, Carlo Ratti, and Vincent D Blondel. Urban gravity: a model for inter-city telecommunication flows. Journal of Statistical Mechanics: Theory and Experiment, 2009(07):L07003, 2009.

[20] Renaud Lambiotte, Vincent D Blondel, Cristobald De Kerchove, Etienne Huens, Christophe Prieur, Zbigniew Smoreda, and Paul Van Dooren. Geographical dispersal of mobile communication networks. Physica A: Statistical Mechanics and its Applications, 387(21):5317–5325, 2008.

[21] David Laniado, Yana Volkovich, Salvatore Scellato, Cecilia Mascolo, and Andreas Kaltenbrunner. The impact of geographic distance on online social interactions. Information Systems Frontiers, 20(6):1203–1218, 2018.

[22] Moshe Levy and Jacob Goldenberg. The gravitational law of social interaction. Physica A: Statistical Mechanics and its Applications, 393:418–426, 2014.

[23] David Liben-Nowell, Jasmin Novak, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Geographic routing in social networks. In Proceedings of the National Academy of Sciences, volume 102, pages 11623–11628, 2005.

[24] Jukka-Pekka Onnela, Samuel Arbesman, Marta C González, Albert-László Barabási, and Nicholas A Christakis. Geographic constraints on social network groups. PLoS one, 6(4):e16939, 2011.

[25] Yan Qiao, Kui Wu, and Xinyu Yuan. AutoTomo: Learning-based traffic estimator incorporating network tomography. IEEE/ACM Transactions on Networking, 32(6):4644–4659, 2024.

[26] Dongrun Qin and Zhi Ding. Transport capacity analysis of wireless in-band full duplex ad hoc networks. IEEE Transactions on Communications, 65(3):1303–1318, 2016.

[27] Liang Qin et al. Satformer: Accurate and robust traffic data estimation for satellite networks. In NeurIPS 2024, 2024.

[28] P. J. Rodríguez-Rueda, J. J. Ruiz-Aguilar, J. González-Enrique, and I. Turias. Origin–destination matrix estimation and prediction from socioeconomic variables using automatic feature selection procedure-based machine learning model. Journal of Urban Planning and Development, 147(4), 2021.

[29] Augustin Soule, Anukool Lakhina, Nina Taft, Konstantina Papagiannaki, Kave Salamatian, Antonio Nucci, and Mark Crovella. Traffic matrices: Balancing measurements, inference and modeling. In Proceedings of the ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pages 362–373, 2005.

[30] The Social Shepherd. 53 twitter statistics marketers must know in 2024, 2024. Accessed: 2025-03-11.

[31] Stefan Thurner and Benedikt Fuchs. Physical forces between humans and how humans attract and repel each other based on their social interactions in an online world. PLoS one, 10(7):e0133185, 2015.

[32] U.S. Census Bureau. 2020 gazetteer files — county. U.S. Census Bureau Geography Division, 2020.

[33] U.S. Census Bureau, Population Division. Annual county and county-equivalent population estimates (vintage 2024). Population Estimates Program (PEP); file `co-est2024-alldata.csv`, 2025. We use `POPESTIMATE2020` to align with the Gazetteer vintage.

[34] Krishna V. Veerubhotla and Wilfred G. W. Origin destination traffic matrix prediction in networks using recurrent layer algorithms. In Proc. 6th IEEE Conf. on Information and Communication Technology (CICT), 2022.

[35] Zhiqing Wei, Huici Wu, Xin Yuan, Sai Huang, and Zhiyong Feng. Achievable capacity scaling laws of three-dimensional wireless social networks. IEEE Transactions on Vehicular Technology, 67(3):2671–2685, 2018.

[36] A. G. Wilson. A statistical theory of spatial distribution models. Transportation Research, 1(3):253–269, 1967.

[37] Salomon Wollenstein-Betech, Chuangchuang Sun, Jing Zhang, Christos G. Cassandras, and Ioannis Ch. Paschalidis. Joint data-driven estimation of origin-destination demand and travel latency functions in multi-class transportation networks. Mitsubishi Electric Research Laboratories TR2022-078, 2022.

[38] Hang Yu, Senlai Zhu, Jie Yang, Yuntao Guo, and Tianpei Tang. A bayesian method for dynamic origin–destination demand estimation synthesizing multiple sources of data. Sensors, 21(15):4971, 2021.

[39] Dong Zhao, Adriana-Simona Mihăiță, Yuming Ou, Hanna Grzybowska, and Mingxi Li. Origin–destination matrix estimation for public transport: A multi-modal weighted graph approach. Transportation Research Part C: Emerging Technologies, 165:104694, 2024.

[40] Kai Zhao, Mirco Musolesi, Pan Hui, Weixiong Rao, and Sasu Tarkoma. Explaining the power-law distribution of human mobility through transportation modality decomposition. Scientific Reports, 5:9136, 2015.

[41] Weiping Zheng, Yiyong Li, Minli Hong, Xiaomao Fan, and Gansen Zhao. Flow-by-flow traffic matrix prediction methods: Achieving accurate, adaptable, low cost results. Computer Communications, 194:348–360, 2022.

[42] Yang Zhou, Yan Shi, and Shanzhi Chen. Capacity and delay analysis for large social-aware mobile ad hoc wireless networks. Applied Sciences, 10(5):1719, 2020.