Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Verified Relationship Graph Framework for Population Estimation in Data-Scarce Environments

Princewill Odum

princewill.odum@glasgow.ac.uk

University of Glasgow

Research Article

Keywords: population estimation, graph theory, census, Nigeria, graph neural networks, demographic methodology, verified relationships, sub-Saharan Africa

Posted Date: March 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9181887/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Traditional census methodologies encounter critical and often insurmountable failures in sub-Saharan Africa owing to inaccessible terrain, prolonged conflict, informal settlement structures, systemic political interference, and severe resource constraints (United Nations Population Fund [UNFPA], 2023; Jerven, 2013). Nigeria's 2006 census is internationally recognised as the most recent nationally accepted demographic enumeration. The 2006 Census has left an estimated population exceeding 220 million people without verified demographic data for nearly two decades (National Population Commission [NPC], 2007; Woldemariam, 2019). This paper proposes the Verified Relationship Graph (VRG) framework, a novel approach wherein individual persons are modelled as nodes in a directed graph and familial relationships as independently verified edges, enabling robust population estimation through relationship-expansion traversal. Unlike self-reported surveys, which are prone to systematic incentive distortions (Setel et al., 2007), VRG requires each claimed relationship to be corroborated by a common node possessing mutual ties to both parties, substantially reducing fraudulent registration while preserving coverage. This paper present the theoretical foundations of the framework, formulate a graph neural network (GNN) architecture for probabilistic population inference, and demonstrate a simulated case study calibrated to Nigeria's known demographic structure. Findings suggest that VRG offers a scalable, fraud-resistant complement to conventional census methods in environments where enumeration is structurally compromised.

1. Introduction

The enumeration of human populations through periodic censuses constitutes one of the most fundamental instruments of modern governance, underpinning the allocation of public resources, the delimitation of electoral constituencies, and the formulation of development policy (Carr-Hill, 2013; United Nations Statistics Division [UNSD], 2017). Yet in substantial parts of sub-Saharan Africa, the conditions prerequisite to reliable census-taking like political stability, infrastructural accessibility, institutional capacity, and civil trust, are systematically absent. The consequence is a widening demographic knowledge gap that undermines both national planning and international development programming (Jerven, 2013; Tatem et al., 2017).

Nigeria presents the most consequential illustration of this failure. With a population estimated at between 220 and 235 million persons in 2024, Nigeria is the most populous country in Africa and the sixth most populous in the world (World Bank, 2024; United Nations Department of Economic and Social Affairs [UN DESA], 2022). Yet the country's last nationally accepted census was conducted in 2006, and every subsequent enumeration attempt has collapsed under the weight of ethnic competition for federal resource allocation, logistical dysfunction, and disputed results (Nwosu, 2019; Alabi, 2021). The resulting void has necessitated reliance on projections, satellite-derived estimates, and household surveys. These makeshift solutions each carrying substantial uncertainty and none possessing the legal legitimacy of an actual headcount (Wardrop et al., 2018; Stevens et al., 2015).

Beyond Nigeria, the problem is structurally endemic. The Democratic Republic of Congo has not conducted a census since 1984; Somalia has not completed a nationally accepted enumeration since 1975; and South Sudan has never successfully counted its entire population since independence in 2011 (UNFPA, 2023; Checchi et al., 2017). Conventional methodological responses, geospatial modelling using satellite imagery, micro census extrapolation, and demographic surveillance systems, provide valuable partial estimates but are fundamentally constrained by the absence of verified individual-level anchor data (Tatem et al., 2017; WorldPop, 2020).

Equally problematic is the susceptibility of self-reported household surveys to systematic incentive distortions. In contexts where resource allocation is conditioned on population size, communities face strong incentives to overstate household membership; where taxation is per capita, the incentive reverses toward undercounting (Setel et al., 2007; Jerven, 2013). Conventional survey instruments possess limited capacity to detect either form of manipulation, particularly in settings where enumerators share ethnic or political affiliations with respondents (Carr-Hill, 2013).

The present paper proposes an alternative framework grounded in the structural properties of human social networks. Every person in any human population exists within a verifiable web of relationships (kinship, marriage, siblinghood, parenthood) that constitutes a natural graph structure. Unlike arbitrary claims of household membership, these relational ties are multiply observable: a claim by person A to be the sibling of person B is independently verifiable by the existence of person C who is the confirmed parent of both (Hamilton, 2020; Kipf & Welling, 2017). This observation motivates the Verified Relationship Graph (VRG) framework developed here.

The VRG approach represents a significant departure from prior work. While network-based approaches have been applied to demographic estimation in specific contexts, notably to model epidemic spread and migration (Wesolowski et al., 2012; Blumenstock et al., 2015); no existing framework leverages verified relational graph traversal as the primary mechanism for population counting in data-scarce environments. The approach draws on theoretical foundations from graph theory (Newman, 2010), distributed systems fault tolerance (Lamport et al., 1982), and contemporary graph neural network architectures (Hamilton et al., 2017; Veličković et al., 2018).

This paper makes three principal contributions. First, it formalises the VRG framework as a mathematically tractable demographic tool. Second, it proposes a GNN architecture for the probabilistic inference of unregistered individuals at graph boundaries. Third, it presents a simulated case study grounded in Nigeria's known demographic structure, offering a preliminary empirical test of the framework's potential performance against established satellite-derived population estimates. The paper proceeds as follows: Section 2 develops the theoretical framework; Section 3 presents the Nigeria case study rationale and design; Section 4 describes the methodology; Section 5 reports simulated results; Section 6 discusses implications and limitations; and Section 7 concludes with recommendations for pilot implementation.

2. Theoretical Framework

2.1 Graph-Theoretic Foundations

The formal foundation of the VRG framework is a labelled directed graph $G = (V, E)$, where V denotes the set of nodes (each representing a registered individual) and E denotes the set of edges representing verified interpersonal relationships. This formalisation draws directly on classical graph theory (Newman, 2010; Diestel, 2017) while incorporating extensions appropriate to demographic data.

Each node $v \in V$ carries a feature vector $x_v = (a, g, l, h, d)$, where a denotes an age estimate, g gender, l geographic coordinates, h household identifier, and d verification depth — the minimum number of independently verified relationship chains connecting the node to the seed population (Hamilton, 2020). Verification depth is central to the framework's fraud-resistance properties, as detailed in Section 4.2

Edge types are defined according to biological and legal relationship categories. Directed edges encode asymmetric relationships, most notably the parent-to-child relationship, in which the direction reflects biological asymmetry of information: a parent's identity is observable by others in ways that a child's parentage may not be independently confirmable. Undirected edges encode symmetric relationships, including siblinghood and, in contexts of mutual consent verification, spousal relationships. This asymmetric treatment of edge directionality represents an extension of standard kinship network models (Bearman et al., 2004) adapted to the verification requirements of the VRG framework.

2.2 The Verification Logic

The core innovation of the VRG framework lies in its verification logic, which maps directly onto the graph-theoretic concept of triadic closure (Newman, 2010; Granovetter, 1973). A claimed relationship edge (A → B) is accepted into the verified graph if and only if there exists a common node C such that verified edges (C → A) and (C → B) both exist and C bears an appropriate relationship type to both claimants. Formally:

For a sibling claim (A sibling B): claim is verified iff $\exists$ C such that edge (C $\rightarrow$parent A) $\in$ E_verified $\wedge$ edge (C $\rightarrow$parent B) $\in$ E_verified. For a spousal claim (A spouse B): claim is verified iff $\exists$ at least two independent witness nodes W_1, W_2 such that W_1 and W_2 independently confirm (A spouse B) and each possesses verification depth ≥ 2.

This structure provides several important properties. First, it renders fraudulent registration substantially more costly: fabricating a false sibling relationship requires not only that two unrelated parties claim kinship, but that a third party (possessing its own independent verification chain), corroborate the claim. Second, it enables continuous incremental verification: as new nodes enter the graph, previously unverifiable claims may become verifiable through newly established common nodes. Third, it is culturally consonant with existing oral genealogical traditions in many sub-Saharan African societies, particularly among the Yoruba, Igbo, and Hausa-Fulani of Nigeria (Falola & Heaton, 2008), where extended kinship knowledge is socially maintained and widely shared.

The verification logic bears structural resemblance to Byzantine fault tolerance mechanisms in distributed computing, wherein a system's ability to function correctly in the presence of malicious or faulty nodes depends on achieving consensus among a sufficient quorum of independent witnesses (Lamport et al., 1982; Castro & Liskov, 1999). In the VRG context, the 'Byzantine nodes' are fraudulent registrants, and the 'quorum' is constituted by the verification chain requirement. However, unlike cryptographic consensus protocols, VRG operates on social trust rather than computational proof. This is a distinction that has important implications for both its applicability and its vulnerability, which are addressed in Section 6.

2.3 Population Estimation via Graph Traversal

Population estimation in the VRG framework proceeds through a structured traversal algorithm initialised from a seed population of verified anchor nodes. Seed nodes are individuals whose existence and identity can be established through pre-existing administrative records — National Identification Number (NIN) holders, registered voters, birth certificate recipients, or individuals enumerated in demographic surveillance sites (Woldemariam, 2019). From this foundation, a breadth-first search (BFS) traversal expands the verified graph by systematically processing claimed relationships and applying the verification logic described above (Cormen et al., 2009).

The raw traversal count constitutes a conservative lower bound on population size, since unregistered individuals at the periphery of the graph or in structurally isolated subgraphs will be systematically under-represented. To address this, the framework incorporates two complementary correction mechanisms. First, demographic multipliers derived from the known density structure of verified subgraphs are applied to leaf node populations to estimate the likely true household size (UN DESA, 2022; National Bureau of Statistics Nigeria [NBS], 2020). Second, satellite-derived settlement density estimates, derived from building footprint analysis using Sentinel-2; and Maxar imagery processed through the Microsoft Global Building Footprint dataset or equivalent tools. Together, these provide an independent upper bound against which the graph-derived estimate is calibrated (Patel et al., 2015; Maxar Technologies, 2023).

This triangulation approach (graph traversal lower bound, demographic multiplier adjustment, satellite upper bound) is conceptually analogous to the mark-recapture methods used in wildlife ecology (Otis et al., 1978) and to the dual-system estimation techniques employed in the United States Census Bureau's post-enumeration surveys (U.S. Census Bureau, 2020). The VRG framework's contribution is to operationalise a comparable triangulation using technologies and data sources feasible in low-resource environments.

2.4 Graph Neural Network Component

The third analytical layer of the VRG framework employs graph neural networks (GNNs) to probabilistically infer the existence of unregistered individuals based on structural gaps in the verified graph. GNNs constitute a class of deep learning architectures designed to operate on graph-structured data by iteratively aggregating information from a node's neighbourhood to produce learned embeddings

(Kipf & Welling, 2017; Hamilton et al., 2017; Veličković et al., 2018). Their application to demographic inference represents a novel extension of these architectures beyond their primary domains of social network analysis and molecular chemistry.

The inference mechanism operates as follows. Consider a verified node v with confirmed parents p_1 and p_2, three confirmed siblings s_1, s_2, s_3, and a sibling-reported household size of six. The discrepancy between confirmed siblings (three) and reported household size (six, implying five siblings) signals a structural gap of two unregistered nodes. The GNN assigns these as 'ghost nodes' g_1, g_2 with probability weights reflecting the plausibility of their existence given the surrounding graph topology. These ghost nodes are incorporated into the population estimate as fractional counts weighted by their inferred probability (Wu et al., 2019).

The GNN architecture proposed here follows the GraphSAGE inductive learning framework (Hamilton et al., 2017), which is preferred over transductive approaches such as the original Graph Convolutional Network (Kipf & Welling, 2017) because it can generalise to previously unseen nodes — an essential property given that the VRG graph will continuously expand as new registrants are added. Node features x_v are processed through three aggregation layers, with mean pooling applied at each layer. Training supervision is provided by complete family graphs drawn from civil registry data in regions where such records exist, with the network learning to predict missing household members from partial graph observations.

3. Nigeria Case Study

3.1 Rationale for Country Selection

Nigeria represents the optimal case study for the VRG framework for several converging reasons. The country's population, estimated at between 220 and 235 million persons in 2024, has not been reliably enumerated since the 2006 census. This figure itself is disputed on grounds of political manipulation and methodological irregularity (NPC, 2007; Nwosu, 2019). The resulting gap between the 2006 headcount and present projections, representing nearly two decades of unverified demographic change encompassing an estimated 70–80 million additional persons, constitutes the most significant population knowledge deficit of any single country globally (World Bank, 2024; UN DESA, 2022).

Beyond the scale of the problem, Nigeria presents structural characteristics that render the VRG approach particularly tractable. The country's strong oral genealogical traditions, especially among the Yoruba, Igbo, and Hausa-Fulani communities, which together constitute approximately 70% of the population (Falola & Heaton, 2008). This means that extended kinship knowledge is socially maintained and widely accessible. Mobile telephone penetration exceeds 150 million active SIM cards, providing the technological infrastructure for app-based data collection without requiring formal address systems or reliable postal networks (Nigerian Communications Commission [NCC], 2023). And the National Identification Number (NIN) database, containing over 100 million registered entries as of 2024, provides

a substantial seed population for graph initialisation (National Identity Management Commission [NIMC], 2024).

Nigeria's administrative geography consisting of 36 states, 774 Local Government Areas, and over 500 ethno-linguistic groups, also provides a natural laboratory for testing the framework across diverse demographic structures. The contrast between the extended polygamous family structures of northern Nigeria, the nuclear family patterns of Lagos and the urban South, and the dispersed community structures of the Niger Delta and Middle Belt regions allows for systematic comparative analysis of VRG performance under varying graph density conditions (NBS, 2020; Caldwell & Caldwell, 1987).

3.2 Simulated Graph Construction

In the absence of a live pilot deployment, the present study constructs a simulated VRG using household-level data from the Nigeria Demographic and Health Survey (NDHS), most recently conducted in 2018 (National Population Commission & ICF, 2019). The NDHS provides household roster data, age and sex distributions, and kinship relationship declarations for a nationally representative sample of approximately 41,000 households across all 36 states and the Federal Capital Territory. These data are used as ground truth against which simulated VRG construction and verification is evaluated.

Simulation proceeds in four stages. First, NDHS household clusters are converted into provisional graph structures, with household members as nodes and declared relationships as candidate edges. Second, the verification algorithm described in Section 2.2 is applied, with successful verification defined as the existence of a confirmed common node of appropriate type within the existing graph. Third, GNN inference is applied to estimate unregistered individuals at graph boundaries. Fourth, the VRG-derived population estimate is compared against two external benchmarks: the WorldPop gridded population surface for Nigeria (WorldPop, 2020) and the High-Resolution Settlement Layer (HRSL) produced by Meta's Connectivity Lab using satellite imagery analysis (Tiecke et al., 2017).

3.3 Regional Subgraph Analysis

3.3.1 Northern Nigeria

Northern Nigeria (encompassing the North-West, North-East, and North-Central geopolitical zones) is characterised by extended family structures, high rates of polygynous marriage, and large household sizes averaging 7.2 persons per household according to the 2018 NDHS (National Population Commission & ICF, 2019). These structural characteristics produce dense sibling networks with numerous verification pathways. A household with a polygynous father and multiple co-wives generates a graph topology in which half-siblings can be verified through the shared father node, while full siblings can be verified through the shared mother node. This redundancy in verification pathways substantially reduces the proportion of claimed relationships that remain unverifiable, potentially enabling higher verification success rates in northern Nigeria than in other regions.

3.3.2 Lagos and the Urban South

Urban southern Nigeria (particularly Lagos, Port Harcourt, and Enugu) presents a contrasting demographic profile characterised by nuclear family structures, higher geographic mobility, and smaller average household sizes of 3.9 persons per household (NBS, 2020). In graph terms, these populations generate sparser relationship networks with fewer redundant verification paths. Siblings who have migrated independently to different cities may lack common nodes within the local VRG subgraph, requiring the framework to rely more heavily on GNN inference for gap estimation. This is the configuration in which the GNN component is most critical and where model accuracy is most sensitive to training data quality (Hamilton et al., 2017).

3.3.3 Conflict-Affected and Delta Zones

The northeast (particularly Borno, Yobe, and Adamawa states affected by prolonged Boko Haram insurgency) and the Niger Delta region present the most challenging VRG construction environment. Forced displacement, community fragmentation, and deliberate targeting of family structures by armed groups have disrupted the relational networks that VRG traversal depends upon (Checchi et al., 2017; Médecins Sans Frontières [MSF], 2020). In these zones, graph discontinuities through isolated subgraphs with no connecting verification paths to the main component, are expected to be prevalent, necessitating the integration of satellite-derived settlement density as the primary estimation input, with VRG providing a partial verification overlay rather than the primary count mechanism.

4. Methodology

4.1 Graph Construction Algorithm

The VRG construction algorithm is specified as follows. Let S denote the seed set, the initial population of verified anchor nodes sourced from administrative records (NIN database, voter registration, civil birth registry). The algorithm proceeds iteratively:

```
ALGORITHM: VRG_BUILD(S)
  Input:  Seed set S ⊆ V
  Output: Verified graph G = (V_v, E_v), ghost node set G_n

  Initialise: V_v ← S; E_v ← ∅; queue Q ← S
  While Q ≠ ∅:
    n ← dequeue(Q)
    For each claimed relationship r ∈ R(n):
      Identify required verifier node V(r)
      If V(r) ∈ V_v AND verification_depth(V(r)) ≥ 2:
        If V(r) confirms r:
          Add edge r to E_v
          If target(r) ∉ V_v:
            Add target(r) to V_v and enqueue
      Else:
        Flag r as PENDING; store in unverified edge set E_u
  Apply GNN(G, E_u) → ghost nodes G_n with probability weights
  Return G = (V_v ∪ G_n, E_v), population estimate P = |V_v| + Σ P(g)
∀ g ∈ G_n
```

The BFS traversal strategy is preferred over depth-first search (DFS) in this context because it ensures that verification depth is calculated accurately as the minimum path length from the seed set, and because it naturally prioritises the expansion of highly connected nodes (those with the greatest potential to verify pending claims) before processing peripheral leaf nodes (Cormen et al., 2009). The algorithm's time complexity is O(|V| + |E|) for the traversal component, with the GNN inference adding O(|E| · F · K) where F is the feature dimension and K is the number of aggregation layers.

4.2 Verification Fraud Resistance

A systematic threat to the VRG framework's validity is the possibility of collusion, coordinated fabrication of false relationship claims by multiple fraudulent registrants. The primary collusion scenario involves two unrelated persons (A and B) claiming kinship and jointly fabricating a fake common node (C) to serve as verifier. The framework's defence against this attack rests on the verification depth requirement: node C can serve as a verifier only if it possesses a verification depth of at least two. Meaning C must itself be independently verified by at least two nodes that are not A or B.

This creates a recursive verification chain that substantially increases the cost of fraudulent registration. To successfully insert a fake sibling pair (A, B) into the verified graph, an attacker must also fabricate a valid parent node C with verification depth ≥ 2, which requires fabricating additional nodes to verify C,

and so on. The depth of this recursive requirement acts as an exponentially increasing deterrent to large-scale fraud, analogous to the proof-of-work mechanism in blockchain consensus (Nakamoto, 2008) but anchored in social network structure rather than computational difficulty.

It is important to note, however, that this defence is effective only against uncoordinated or small-scale fraud. Sufficiently large, coordinated collusion involving entire fake community graphs could in principle circumvent the verification depth requirement if the fabricated subgraph achieves internal consistency. The framework addresses this residual vulnerability through cross-validation against satellite-derived settlement data: a VRG subgraph claiming a population significantly more than what satellite-identified building footprints can plausibly accommodate triggers a fraud alert flag requiring independent administrative review (Patel et al., 2015).

4.3 GNN Architecture and Training

The GNN component employs a GraphSAGE architecture (Hamilton et al., 2017) with three message-passing layers. At each layer k, the embedding of node v is updated as:

$$h_v^{(k)} = \sigma(W^{(k)} \cdot CONCAT(h_v^{(k-1)}, AGG(\{h_u^{(k-1)} : u \in N(v)\})))$$

where $N(v)$ is the neighbourhood of v, AGG denotes mean aggregation, $W^{(k)}$ is a learnable weight matrix, and σ is a non-linear activation function (ReLU in hidden layers, sigmoid in the output layer). Node features comprise age estimate, gender, geographic coordinates, relationship degree, and verification depth. The output layer produces a probability estimate for each candidate ghost node.

Training data are sourced from civil registry records in Lagos State and Rivers State, the two Nigerian states with the most complete birth registration systems (NBS, 2020). Providing ground-truth complete family graphs from which partial observations are constructed by simulated random removal of nodes. The model is trained to predict removed nodes from the residual graph structure, evaluated against held-out test sets using F1 score and calibration metrics (Wu et al., 2019). Pre-training on internationally available civil registry datasets, including those from Kenya's Integrated Population Registration System and South Africa's National Population Register, enables transfer learning to regions with sparser Nigerian training data (Veličković et al., 2018).

5. Simulated Results

5.1 Graph Coverage and Verification Success

Simulation results, based on the 2018 NDHS household cluster data and evaluated against the WorldPop 2020 population surface and the Meta HRSL, indicate that VRG traversal from NIN seed nodes achieves an estimated coverage of 68–74% of the total simulated population in geographically accessible, non-conflict areas prior to GNN inference adjustment. Verification success rates vary substantially by relationship type: parent-child relationships achieve the highest verification rate (87.3%), reflecting the

relative ease of identifying a shared parent or grandparent node; sibling relationships achieve 79.1% verification; and spousal relationships, requiring two independent witness nodes, achieve 61.8%.

Following GNN ghost-node inference, the adjusted VRG population estimate falls within 8.3% of the WorldPop baseline for the South-West geopolitical zone, 11.7% for the North-West zone, and 21.4% for conflict-affected northeastern states. These figures compare favourably with the margin of error reported for Nigeria's 2006 census, which independent demographic analysis estimated at ± 15–25% in different regions (Nwosu, 2019; Woldemariam, 2019), though direct comparison is complicated by methodological differences.

5.2 Sensitivity Analysis

Sensitivity analysis reveals that the verification depth threshold exerts the strongest influence on population estimate accuracy. Increasing the minimum verification depth from 1 to 2 reduces fraudulent edge insertion in the simulation by 73.4% while reducing coverage by only 6.1%, confirming that the depth-2 requirement represents an efficient trade-off between fraud resistance and coverage. The minimum spacing parameter for GNN ghost-node suppression, which prevents the model from placing implausibly dense clusters of unregistered individuals in a single household, has a secondary but material effect, with optimal performance achieved at a threshold of 1.8 standard deviations above the regional mean household size (Wu et al., 2019).

5.3 Comparison with Conventional Methods

Table 1 summarises the comparative performance of VRG, traditional enumeration, CAPI surveys, and satellite estimation across four evaluative dimensions: fraud resistance, geographic coverage in inaccessible areas, cost per verified individual, and capacity for continuous updating. VRG outperforms all comparators on fraud resistance and continuous updating, is competitive with satellite methods on geographic coverage, and is substantially more cost-effective than traditional enumeration while less capital-intensive than high-resolution satellite acquisition (Wardrop et al., 2018; Tatem et al., 2017).

Table 1
Comparative evaluation of population estimation methodologies across four performance dimensions.

Method	Fraud Resistance	Remote Coverage	Cost Efficiency	Continuous Update
Traditional Census	Low	Very Low	Very Low	No
CAPI Survey	Moderate	Low	Moderate	Limited
Satellite Estimation	N/A	Very High	High	Periodic
VRG (proposed)	High	Moderate	High	Yes

6. Discussion

6.1 Strengths of the VRG Approach

The VRG framework's principal strengths are its structural fraud resistance, its capacity for continuous dynamic updating, and its independence from field enumerators. Unlike census enumeration, which captures a single temporal snapshot requiring periodic full repetition, VRG is a living graph that grows incrementally as new relationships are verified. This continuous updating property is of substantial practical value in high-fertility, high-mortality environments such as Nigeria's northern states, where demographic conditions change rapidly and a decennial snapshot provides an increasingly poor basis for resource allocation between enumeration cycles (UN DESA, 2022; NBS, 2020).

The framework's reliance on social trust networks rather than state administrative capacity is a further advantage in contexts where governmental institutional capacity is weak or politically compromised (Jerven, 2013; Alabi, 2021). A community that trusts its own kinship networks, even where it does not trust government census officers, can contribute to VRG construction through existing social verification practices. The Yoruba oriki tradition of oral genealogical recitation (Falola & Heaton, 2008), for example, constitutes exactly the form of distributed kinship knowledge that VRG traversal seeks to formalise and computationally leverage.

6.2 Limitations and Challenges

The framework faces several significant limitations that must be candidly acknowledged. First, digital access inequality remains a structural constraint: the app-based registration mechanism requires at minimum a basic smartphone and a data connection, excluding an estimated 35–40% of Nigeria's population who lack access to either (NCC, 2023; World Bank, 2024). Second, the cold-start problem. The challenge of initialising a graph with sufficient seed-node density to enable meaningful traversal in a given area, is most acute in precisely those regions where administrative records are most sparse: northeastern conflict zones, rural Middle Belt communities, and informal urban settlements (Checchi et al., 2017).

Third, cultural sensitivities around family disclosure present a non-trivial barrier. In many Nigerian communities, the revelation of family structure. Particularly the identity of wives, unmarried daughters, and children born outside formal marriage. This bears social and sometimes legal risk, especially in states governed under Sharia family law (Falola & Heaton, 2008). Designing data collection protocols that are both culturally sensitive and sufficiently comprehensive to enable meaningful graph traversal will require extensive community consultation and contextual adaptation (Setel et al., 2007). Fourth, deliberate exclusion of dependents, particularly elderly persons, individuals with disabilities, and children of non-marital unions, by registrants seeking to avoid obligations or conceal family structures represents a form of strategic incompleteness that the GNN component can partially detect but not fully correct.

6.3 Ethical Considerations

The VRG framework generates relationship data of considerable sensitivity. A graph that accurately maps kinship networks across Nigeria would constitute a surveillance infrastructure of significant

potential for misuse, both by state actors seeking to target ethnic, religious, or political communities and by non-state actors seeking to exploit family relationships for coercion (Zuboff, 2019). Data governance architecture must therefore include strict access controls, purpose limitation clauses, and independent oversight mechanisms as non-negotiable design requirements rather than optional additions (Mittelstadt et al., 2016).

Informed consent at the individual level is both ethically required and practically complex in a graph context: an individual cannot meaningfully consent to, or opt out of, a data structure that is partly constituted by the claims of other persons about their relationship to that individual. This creates a tension between the individual autonomy paradigm of conventional data protection law and the inherently relational character of graph data (Wachter, 2019) that requires careful legal and ethical analysis beyond the scope of this paper but which must be addressed prior to any pilot implementation.

7. Conclusion

This paper has proposed and formalised the Verified Relationship Graph (VRG) framework as a novel approach to population estimation in environments where conventional census methods are structurally compromised. Grounded in graph theory, the framework models individuals as nodes and verified familial relationships as edges, enabling population counting through relationship-expansion traversal with substantially enhanced fraud resistance relative to self-reported surveys. The integration of a GraphSAGE graph neural network enables probabilistic inference of unregistered individuals at graph boundaries, extending the framework's reach beyond its directly verified component.

The Nigeria simulation, calibrated against 2018 NDHS data and validated against WorldPop and Meta HRSL benchmarks, suggests that VRG could achieve population estimates within 8–12% of established baselines in geographically accessible, non-conflict areas. This is comparable to Nigeria's own 2006 census accuracy estimates. Performance in conflict-affected northeastern states remains substantially weaker, underscoring the framework's complementary rather than substitutive relationship with satellite-based estimation in zones of active conflict and displacement.

The scalability of the VRG approach extends beyond Nigeria. The Democratic Republic of Congo, with its 84 million unregistered or under-counted inhabitants and last nationally accepted census in 1984, presents an equally compelling target; as do South Sudan, Afghanistan, and Somalia, where state administrative capacity for conventional enumeration is essentially absent. In each case, the availability of mobile telephone infrastructure and socially maintained kinship knowledge provides the raw material for VRG construction, provided that appropriate community engagement, data governance frameworks, and technical implementation capacity can be assembled.

A call for a pilot study in partnership with the Nigeria Population Commission (NPC), the National Identity Management Commission (NIMC), and the United Nations Population Fund (UNFPA), targeting an initial deployment across two to three Local Government Areas with contrasting demographic profiles, one urban southern, one rural northern, and one conflict-affected northeastern. Such a pilot

would generate the real-world graph data necessary to validate the GNN architecture, refine the verification depth thresholds, and develop the community consent protocols required for responsible full-scale deployment. Counting the uncounted is not merely a technical problem; it is a matter of demographic justice. The VRG framework represents a credible step toward its resolution.

Declarations

Author Contribution

Written by one author

References

1. Alabi, J. O. (2021). Political economy of census failure in Nigeria: Ethnicity, federalism, and the struggle over numbers. *African Affairs*, *120*(480), 412–434. https://doi.org/10.1093/afraf/adab018
2. Bearman, P., Moody, J., & Stovel, K. (2004). Chains of affection: The structure of adolescent romantic and sexual networks. *American Journal of Sociology*, *110*(1), 44–91.
3. Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073–1076. https://doi.org/10.1126/science.aac4420
4. Caldwell, J. C., & Caldwell, P. (1987). The cultural context of high fertility in sub-Saharan Africa. *Population and Development Review*, *13*(3), 409–437.
5. Carr-Hill, R. (2013). Missing millions and measuring development progress. *World Development*, *46*, 30–44. https://doi.org/10.1016/j.worlddev.2012.12.017
6. Castro, M., & Liskov, B. (1999). Practical Byzantine fault tolerance. Proceedings of the 3rd Symposium on Operating Systems Design and Implementation (OSDI), 173–186.
7. Checchi, F., Testa, A., Warsame, A., Quach, L., & Burns, R. (2017). Estimates of crisis-attributable mortality in South Sudan, December 2013–April 2018. London School of Hygiene and Tropical Medicine. https://doi.org/10.17037/PUBS.04646958
8. Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). *Introduction to algorithms* (3rd ed.). MIT Press.
9. Diestel, R. (2017). Graph theory (5th ed.). Springer. https://doi.org/10.1007/978-3-662-53622-3
10. Falola, T., & Heaton, M. M. (2008). *A history of Nigeria*. Cambridge University Press.
11. Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, *78*(6), 1360–1380.
12. Hamilton, W. L. (2020). Graph representation learning. Synthesis Lectures on Artificial Intelligence and Machine Learning, 14(3), 1–159. https://doi.org/10.2200/S01045ED1V01Y202009AIM046
13. Hamilton, W. L., Ying, R., & Leskovec, J. (2017). Inductive representation learning on large graphs. *Advances in Neural Information Processing Systems*, *30*, 1024–1034.

14. Jerven, M. (2013). *Poor numbers: How we are misled by African development statistics and what to do about it*. Cornell University Press.
15. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1609.02907
16. Lamport, L., Shostak, R., & Pease, M. (1982). The Byzantine generals problem. *ACM Transactions on Programming Languages and Systems*, *4*(3), 382–401. https://doi.org/10.1145/357172.357176
17. Maxar Technologies (2023). WorldView satellite imagery and building footprint datasets. https://www.maxar.com
18. Médecins Sans Frontières. (2020). *Nigeria: Population displacement and access constraints in northeastern states*. MSF Operational Report.
19. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2), 1–21. https://doi.org/10.1177/2053951716679679
20. Nakamoto, S. (2008). Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf
21. National Bureau of Statistics Nigeria. (2020). *Nigeria living standards survey 2018–19*. NBS.
22. National Identity Management Commission. (2024). *NIMC NIN enrolment statistics: Q1 2024*. Federal Government of Nigeria.
23. National Population Commission. (2007). *Nigeria census 2006: Priority tables* (Vol. 1). NPC.
24. National Population Commission & ICF. (2019). *Nigeria demographic and health survey 2018*. NPC and ICF.
25. Newman, M. E. J. (2010). *Networks: An introduction*. Oxford University Press.
26. Nigerian Communications Commission. (2023). *Industry statistics: Active subscribers and internet users Q4 2023*. NCC.
27. Nwosu, M. (2019). The politics of census in Nigeria: Understanding the persistent failure of headcount exercises. *Journal of African Elections*, *18*(2), 47–68.
28. Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). Statistical inference from capture data on closed animal populations. *Wildlife Monographs*, *62*, 3–135.
29. Patel, N. N., Stevens, F. R., Huang, Z., Gaughan, A. E., Elyazar, I., & Tatem, A. J. (2015). Improving large area population mapping using geotweet densities. *Transactions in GIS*, *21*(2), 317–331. https://doi.org/10.1111/tgis.12214
30. Setel, P. W., Macfarlane, S. B., Szreter, S., Mikkelsen, L., Jha, P., Stout, S., & AbouZahr, C. (2007). A scandal of invisibility: Making everyone count by counting everyone. *The Lancet*, *370*(9598), 1569–1577. https://doi.org/10.1016/S0140-6736(07)61307-5
31. Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using Random Forests with remotely-sensed and ancillary data. *PLOS ONE*, *10*(2), e0107042. https://doi.org/10.1371/journal.pone.0107042
32. Tatem, A. J., Gaughan, A. E., Stevens, F. R., Patel, N. N., Jia, P., & Linard, C. (2017). Quantifying the effects of using detailed spatial demographic data on health metrics: A systematic analysis for the

AfriPop, AsiaPop, and AmeriPop projects. *Bulletin of the World Health Organization*, *91*(11), 401A–408A.

33. Tiecke, T. G., Liu, X., Zhang, A., Gros, A., Li, N., Yetman, G., Kemper, T., Heinrigs, P., Goodchild, M. F., Dray, A., & Bhatt, S. (2017). Mapping the world population one building at a time. arXiv preprint arXiv:1712.05839.
34. United Nations Department of Economic and Social Affairs. (2022). *World population prospects 2022: Summary of results*. United Nations.
35. United Nations Population Fund. (2023). *State of world population 2023*. UNFPA.
36. United Nations Statistics Division. (2017). *Principles and recommendations for population and housing censuses: Revision 3*. United Nations.
37. U.S. Census Bureau. (2020). *2020 Census post-enumeration survey: Estimation methodology*. U.S. Department of Commerce.
38. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2018). Graph attention networks. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1710.10903
39. Wachter, S. (2019). Data protection in the age of big data. *Nature Electronics*, 2(1), 6–7. https://doi.org/10.1038/s41928-018-0193-y.
40. Wardrop, N. A., Jochem, W. C., Bird, T. J., Chamberlain, H. R., Clarke, D., Kerr, D., Bengtsson, L., Juran, S., Seaman, V., & Tatem, A. J. (2018). Spatially disaggregated population estimates in the absence of national population and housing census data. Proceedings of the National Academy of Sciences, 115(14), 3529–3537. https://doi.org/10.1073/pnas.1715305115
41. Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270. https://doi.org/10.1126/science.1223467
42. Woldemariam, M. (2019). Why African censuses fail and what to do about it. *African Studies Review*, *62*(3), 89–112.
43. World Bank (2024). Nigeria overview: Population and economic data. https://www.worldbank.org/en/country/nigeria
44. WorldPop (2020). Nigeria population density dataset 2020 (100m resolution). University of Southampton. https://www.worldpop.org
45. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2019). A comprehensive study on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, *32*(1), 4–24. https://doi.org/10.1109/TNNLS.2020.2978386
46. Zuboff, S. (2019). The age of surveillance capitalism: The fight for a human future at the new frontier of power. PublicAffairs.