A Graph-Theoretic Feasibility Analysis of Elite Mate Search in China's Tier-One Cities

ROMANCE Working Paper No. 001 · calibrated synthetic market study

Prepared for the ROMANCE Research Institute · synthetic network-and-matching experiment calibrated to public demographic and wage statistics

ABSTRACT

This paper is anchored to public 2024 population, wage, and economic data for Beijing, Shanghai, and Shenzhen, together with 2020 census evidence on sex ratios, migration, and education, and public marriage-timing statistics for urban China and Shanghai.[2]-[15] The target subgroup is defined as men in the three tier-one cities who are graduates of former-985 / top Double-First-Class universities and earn at least RMB 200,000 in annual pre-tax wage income. A BigCLAM-style overlapping-affiliation graph generates latent access to city, education, sector, migrant, age, lifestyle, and affluence circles.[1] A sequential two-sided monthly matching simulator then maps network access into one-year exclusive-match probabilities.

In the calibrated base case, the market contains 4,200 active single agents (1,995 women and 2,205 men), 191,996 realized graph edges, and 43 target men (1.95% of men). Across 1,000 Monte Carlo runs, the average 12-month probability that a woman secures a match to the target subgroup is 1.45%, while the probability of any exclusive match is 33.33%. Beijing has the highest target-match probability at 2.09%; Shanghai and Shenzhen are both near 1.1%. The strongest result is the access gradient: women in the top elite-access quartile achieve a 3.54% target-match probability, versus 0.33% in the bottom quartile. Women aged 30+ are slightly lower than women aged 25-29 in the base calibration, but the difference is modest relative to supply scarcity and network access. The paper therefore supports a narrow conclusion: in a scarce, high-status urban dating market, the first-order bottleneck is not age alone but *access to the overlapping communities through which scarce partners are actually reachable*.

Keywords: synthetic data, overlapping communities, BigCLAM, matching markets, China, Beijing, Shanghai, Shenzhen, education, salary threshold.

1. Introduction

The original idea was intentionally provocative: use graph theory to estimate whether a woman aged 30+ can still find an elite urban partner in China. As a joke-paper premise, that works immediately. As a machine-learning paper, however, it only becomes defensible if the data-generating process is calibrated to real demographic and labor-market structure rather than to arbitrary assumptions. The present draft therefore treats the problem as a *calibrated synthetic market*, not as a claim to have recovered "true" dating probabilities from real people.

The graph component follows the logic of BIGCLAM: nodes may belong to several circles at once, and ties emerge from shared affiliation strength.[1] That is a natural fit for tier-one-city partner search, where city, education, sector, migrant status, lifestyle, and affluence all overlap. Yet BIGCLAM by itself does not answer the user-facing question. A woman does not merely need a graph edge to a desirable man; she needs exposure, mutual interest, and a successful capacity-constrained match before someone else gets there first. Accordingly, the paper uses a two-stage design: first an overlapping-community graph, then a one-year sequential matching simulation.

The target subgroup is defined conservatively as men in Beijing, Shanghai, or Shenzhen who satisfy two conditions simultaneously: (i) graduate of a former-985 / top Double-First-Class university, and (ii) annual pre-tax wage income of at least RMB 200,000. This phrasing modernizes the legacy "985" shorthand using current Ministry of Education terminology. The Ministry's second-round Double First-Class announcement covers *147 universities*, while the older 985 project ultimately contained 39 institutions.[16][17]

> **Calibration principle.** Every input is tagged as either *directly anchored* to public data or treated as an explicit *scenario assumption*. This distinction matters more than cosmetic precision. The simulation should look like a paper, but it should also be honest.

2. Data anchors and calibration logic

2.1 City size, sex ratio, migration, and broad human-capital structure

The market is split into three city submarkets using 2024 resident-population weights: 21.832 million for Beijing, 24.8026 million for Shanghai, and 17.9895 million for Shenzhen.[2]-[4] These imply synthetic city weights of 33.8%, 38.4%, and 27.8%, respectively. Sex ratios are anchored to the 2020 census: 104.7 in Beijing, 107.33 in Shanghai, and 122.43 in Shenzhen.[5][6][8] The Shenzhen ratio is notably male-heavy and materially changes scarcity once the market is capacity constrained.

Migrant status is also treated as first-order structure rather than a decorative covariate. Public sources report migrant/non-hukou shares of 38.5% in Beijing, 42.1% in Shanghai, and 64.9% in Shenzhen.[4]-[6] This matters because migration is not only a demographic fact; it is also a network fact. A city with a large newcomer population produces more alumni, hometown, and workplace-dependent matching paths, and fewer long-established kinship paths.

For broad education, the simulation uses city-level census college-or-above rates as lower-bound anchors: 41,980 per 100,000 in Beijing, 33,872 per 100,000 in Shanghai, and 28,849 per 100,000 in Shenzhen.[5]-[8] Shanghai's census press briefing adds a stronger working-age anchor: among residents aged 16-59, the college-or-above share was 46.4%.[7] I use that 46.4% figure to scale working-age college-plus targets across the three cities, yielding synthetic active-market targets of 57.5%, 46.4%, and 39.5% for Beijing, Shanghai, and Shenzhen. This step is partly calibrated and partly inferential, but it is anchored to public data rather than invented from scratch.

> **Short quote from the official Shanghai marriage release:**
> "全市初婚平均年龄30.1岁，其中男性30.8岁，女性29.5岁。"[9]

2.2 Marriage timing and the active-singles age distribution

The simulation does not model the whole population; it models active singles aged 25-39. That age distribution should reflect later marriage in large cities. Two public anchors are used. First, the NUS East Asian Institute summary of China's 2020 census places mean first marriage at 29.4 for men and 28.0 for women nationally, and 28.1 for urban women specifically.[10] Second, Shanghai's 2024 administrative marriage data show a substantially later pattern: 30.8 for men and 29.5 for women at first marriage.[9]

I therefore derive base singlehood weights from the public 2020 singlehood rates for ages 25-29 and 30-34, then shift the synthetic tier-one-city market modestly older. For women, the public figures are 33.2% single at ages 25-29 and 9.3% at ages 30-34; for men and women combined the figures are 52.9% and 20.5%, respectively.[10] The resulting active-singles means are about 29.9 in Beijing, 30.1 in Shanghai, and 29.5 in Shenzhen for women, and roughly half a year higher for men. These are *calibrated scenario weights*, not direct observed age distributions for active daters.

2.3 Wage calibration

The wage process is the most consequential part of the data-generating logic because the target subgroup is explicitly defined by income. For Beijing, I anchor the citywide mean to the official 2024 full-caliber urban employment average wage of RMB 143,244.[11] For Shanghai, I anchor to the official 2024 full-caliber urban employment average wage of RMB 12,434 per month, i.e. RMB 149,208 annually.[12] For Shenzhen, the retrieved official sources provide non-private and private averages but not a single citywide full-caliber 2024 value: RMB 174,478 for urban non-private units and RMB 95,217 for urban private units.[13] Accordingly, the synthetic Shenzhen all-unit anchor is set at roughly RMB 142,000, which is an explicit scenario blend of the two official series.

Sectoral wage structure is handled differently across cities. Shenzhen has the strongest directly usable public detail because the 2024 wage-guidance report publishes wage percentiles by industry. For example, the report gives medians / 75th percentiles / 90th percentiles of RMB 194,299 / 351,450 / 607,946 for finance, 167,250 / 256,901 / 372,180 for software and information technology services, and 121,516 / 186,000 / 278,400 for scientific research and technical services.[15] Those percentiles are used to fit log-normal sector distributions. Beijing sector means are anchored to official yearbook wage tables for finance, software and IT, scientific research and technical services, business services, and real estate.[11][14] Shanghai sector means are obtained more conservatively by scaling national non-private sector means to Shanghai's citywide average wage level.[12][14]

The official wage definitions treat these series as *pre-tax wage income*, including bonuses and regular allowances, while excluding capital gains and equity-like returns.[11][13] That is why the paper's RMB 200,000 threshold is interpreted strictly as *pre-tax wage income*, not total wealth or annualized compensation including equity.

2.4 Which assumptions are directly validated, and which are scenario choices?

- **Directly anchored:** city weights, sex ratios, migrant shares, broad college-plus lower bounds, Shanghai working-age college share, official citywide wage means, official Shenzhen wage percentiles, and urban/Shanghai marriage-timing moments.[2]-[15]
- **Scenario assumptions:** the share of former-985 / top-Double-First-Class graduates inside the 25-39 college-educated pool, the exact active-singles age distribution, Shenzhen's blended all-unit wage anchor, and sector assignment probabilities for active singles.

The most important unobserved quantity is the stock share of former-985 / top-Double-First-Class graduates in the current 25-39 dating-age population. Public sources clearly identify the institutional lists but do not directly identify the resulting city-level stock share for the dating pool.[16][17] I therefore treat the top-tier-education share as a conservative scenario parameter inside the college-educated population rather than as a known fact.

3. Synthetic market design

3.1 Population

The calibrated market contains **4,200** active single agents: **1,995** women and **2,205** men. The final synthetic male shares are 51.2% in Beijing, 51.8% in Shanghai, and 55.1% in Shenzhen, closely matching the public anchors.

Each agent receives city, sex, age, migrant status, education tier, sector, two hobby labels, extroversion, and wage income. Education is represented by four tiers: non-degree, college non-elite, mid-elite (211 / non-top Double-First-Class), and top-elite (former-985 / top Double-First-Class). The target male subgroup is then defined as:

$$\text{Target male} = 1[\text{sex} = \text{male}] \times 1[\text{education} = \text{top-tier}] \times 1[\text{salary} \geq 200{,}000 \text{ RMB}]$$

In the realized market this produces **43** target men, or **1.95%** of the male pool. The resulting city counts are 22 in Beijing, 13 in Shanghai, and 8 in Shenzhen.

3.2 Graph model

The graph has **4,200** nodes, **191,996** undirected edges, and mean degree **91.4**. Each node receives a nonnegative affiliation vector over **11** overlapping communities: three city circles, two sector clusters, an elite-education circle, a migrant circle, an age-30-plus circle, two lifestyle circles, and an affluence circle. Edges are drawn from the standard affiliation link function:

$$P(A_{uv} = 1) = 1 - \exp(-F_u^T F_v)$$

This is deliberately BIGCLAM-like.[1] The point is not that romance literally is an affiliation graph, but that access to scarce partners is shaped by overlapping circles: city, education, sector, migration history, and lifestyle are not nested categories; they overlap.

3.3 Matching simulator

The matching layer is a 12-month sequential simulator with one-to-one capacity. Each month, each active woman receives a Poisson number of exposures with a city-specific mean, modulated by extroversion. Candidate men are drawn from a city-aware shortlist whose weights rise with same-city status, graph affinity, direct observed network contact, sector-cluster overlap, and migrant-status similarity. If woman *i* meets man *j*, mutual acceptance is determined by two logistic rules:

$$\text{logit } p^{(w)}_{ij} = \alpha_w + \beta_1 \cdot \text{affinity}_{ij} + \beta_2 \cdot \text{sameCity}_{ij} + \beta_3 \cdot \text{salary}_j + \beta_4 \cdot \text{ageFit}_{ij} + \dots$$

$$\text{logit } p^{(m)}_{ij} = \alpha_m + \gamma_1 \cdot \text{affinity}_{ij} + \gamma_2 \cdot \text{sameCity}_{ij} + \gamma_3 \cdot \text{salary}_i + \gamma_4 \cdot \text{ageFit}_{ij} + \dots$$

A successful mutual acceptance immediately removes both agents from the market. This greedy monthly process is intentionally simple, but it captures the most important scarcity mechanism: a target man can only match once. In the base case, the target subgroup fills at a mean rate of **67.1%** over the 12-month horizon.

4. Validation against public anchors

4.1 City-level macro calibration

Table 1 shows that the simulated city shares, sex ratios, migrant shares, working-age college-plus shares, and citywide wage means align closely with the public anchors used to build the market. This is the part of the model I would call genuinely *validated*.

City	Pop. wt target	Pop. wt sim	Male target	Male sim	Migrant target	Migrant sim	College+ target	College+ sim	Avg wage target (RMB)	Avg wage sim (RMB)
Beijing	33.8%	32.5%	51.1%	51.2%	38.5%	38.5%	57.5%	57.8%	143,244	143,244
Shanghai	38.4%	39.5%	51.8%	51.8%	42.1%	42.1%	46.4%	47.3%	149,208	149,208
Shenzhen	27.8%	28.0%	55.0%	55.1%	64.9%	64.9%	39.5%	41.5%	142,000	142,000

Table 1. City-level calibration targets versus simulated values. The Shenzhen average-wage target is a scenario blend because the retrieved sources provide official private and non-private means but not a single all-unit full-caliber 2024 figure.[13]

4.2 Shenzhen sectoral wage validation

Shenzhen is the strongest test of the wage model because the wage-guidance report publishes industry percentiles rather than only means.[15] Table 2 compares the simulated Shenzhen sector distributions to the official medians, 75th percentiles, and 90th percentiles. The fit is not perfect—especially in the finance tail—but the ordering and magnitude are reasonable, which is enough for a calibrated synthetic model.

sector	median_sim	median_official	p75_sim	p75_official	p90_sim	p90_official
finance	220,345	194,299	389,049	351,450	466,925	607,946
software_it	167,427	167,250	254,611	256,901	411,727	372,180
sci_tech	109,823	121,516	186,383	186,000	244,902	278,400
business_services	94,888	97,392	139,294	158,000	251,468	255,726
real_estate	80,332	81,128	157,769	128,807	245,376	233,085
other	105,215		172,672		252,111	

Table 2. Shenzhen sectoral wage percentiles: official versus simulated. The "other" sector is a residual bucket and has no direct official comparator in this table.

5. Results

5.1 Market composition and scarcity

The realized market contains only **43** target men across **2205** men. That immediately imposes an arithmetic upper bound on overall target-match rates. Even if every target man eventually matches, the market-wide target-match probability for women cannot exceed roughly 2.16% under one-to-one matching. The simulated mean target fill rate of 67.1% therefore already implies that the target subgroup is being used fairly efficiently.

Group	Any match	Target match
All women	33.33%	1.45%
Women age 30+	32.77%	1.38%
Age 25-29	33.67%	1.49%
Age 30-34	34.26%	1.39%
Age 35-39	30.24%	1.36%
Access Q1	27.75%	0.33%
Access Q4	40.94%	3.54%
Status baseline	25.83%	0.55%
Status high	43.84%	4.14%

Table 3. Main outcome summary across 1,000 Monte Carlo runs.

The headline numbers are straightforward. The average woman has a **33.33%** chance of forming any exclusive match within 12 months, but only a **1.45%** chance of matching the target subgroup. Women aged 30+ sit at **1.38%**, only slightly below the 25-29 group in the base calibration. In other words: once supply is genuinely scarce, the massive jump from "can I match anyone?" to "can I match this very specific subgroup?" dominates the age effect.

5.2 City-level results

Beijing has the strongest target-market outcome in the synthetic base case because it combines the highest calibrated working-age college share with the deepest top-tier education stock and the thickest high-status service wage tail. Shanghai and Shenzhen have slightly better general-match probabilities, but the specific target subgroup is thinner there once the education criterion is imposed.

Figure 1. Target-match probability by city. Beijing is highest because the simulated stock of top-tier educated, 200k+ men is thickest there.

Across cities, all-women target-match probabilities are 2.09% in Beijing, 1.15% in Shanghai, and 1.08% in Shenzhen. The 30+ subgroup is similar: 2.08%, 1.04%, and 1.07%, respectively. This is not evidence that age "does not matter"; it is evidence that in a market with only 43 target men, age is not the only or even primary bottleneck.

City	Any match	Target match	Age 30+ target match
Beijing	31.56%	2.09%	2.08%

City	Any match	Target match	Age 30+ target match
Shanghai	33.99%	1.15%	1.04%
Shenzhen	34.56%	1.08%	1.07%

5.3 The access gradient

The strongest result in the paper is the access gradient. I define an "elite-access" score as the probability mass that a woman's exposure shortlist places on target men, which is itself generated by same-city status, graph affinity, sector overlap, and direct network contact. The gradient is steep: women in the top quartile of access achieve a **3.54%** target-match probability, versus only **0.33%** in the bottom quartile—a ratio of roughly **10.8×**. At the decile level, the top decile reaches 5.04% versus 0.26% in the bottom decile.

Figure 2. Target-match probability rises sharply with graph-derived elite access. The nonlinearity in the top deciles is a core result of the model.

This is the paper's main claim. Once the target subgroup becomes thin, access to the right overlapping communities matters much more than broad demographic labels alone. The quantity that matters is not "how many elite men exist in the city" but "how much of their exposure mass lands inside a woman's reachable search frontier."

5.4 Scenario contrast

Figure 3 contrasts two stylized empirical scenarios inside each city: a baseline woman in the bottom access quartile and a high-status woman in the top access quartile. These are not hypothetical hand-coded profiles; they are averages over the simulated subgroups. The gap is dramatic.

Figure 3. Scenario contrast by city. The main difference is not city alone; it is city plus access plus own-status sorting.

In Beijing, the target-match probability rises from 0.26% for the baseline / Q1-access subgroup to 5.60% for the high-status / Q4-access subgroup. The corresponding rise is from 0.32% to 3.27% in Shanghai and from 0.33% to 2.88% in Shenzhen. The humorous implication is that "location, location, location" should be amended to "location, network, and queue position."

5.5 Predictive value of graph access

A five-fold cross-validated ridge model using only baseline observables—age, city, education, migrant status, and salary—achieves an average R^2 of **0.538**. Adding graph-access variables (elite shortlist mass and top affinity to the target subgroup) raises cross-validated R^2 to **0.906**. In this synthetic design, graph access is therefore not a decorative covariate; it is a major sufficient statistic for target-match feasibility.

6. Discussion

Three substantive conclusions survive the recalibration. First, the earlier uncalibrated simulation overstated how casually one can assign education and salary labels in a large-city market. The

present version is stricter: city weights, sex ratios, migrant shares, and broad wage structure are now tied to public sources. Second, after calibration the target subgroup becomes genuinely scarce. That scarcity alone explains why target-match probabilities are small even when overall match probabilities remain reasonable. Third, the access gradient remains strong despite the more realistic macro structure.

What should *not* be concluded? The paper does not show that 30+ women in China "really" have a 1.38% annual chance of finding such a man. That would be a category mistake. The model is calibrated, not observed. It demonstrates that under a public-data-consistent macro structure, the crucial bottleneck is the interaction of scarce supply, network access, and capacity constraints. That is a much narrower and more defensible claim.

7. Limitations

The most important limitation is obvious: the market is synthetic. Several inputs remain scenario choices, especially the stock share of former-985 / top-Double-First-Class graduates in the dating-age pool, the precise active-singles age distribution, and Shenzhen's all-unit wage anchor. The matching rule is also stylized. Real dating markets include family preferences, app design, appearance, personality, religion, hometown, hukou, and strategic timing effects that are not modeled here. In addition, the sequential greedy matcher is intentionally simple and should be read as a capacity-constrained search process, not as a behavioral model of courtship.

There is also an ethical limit. A model like this should be read as a stylized research note or satire-tinged quantitative essay, not as a scoring device for real individuals. The combination of age, education, income, and city can easily drift into profiling if the playful frame is forgotten.

8. Conclusion

After strict recalibration, the original idea still works—but in a more disciplined way. A graph-theoretic treatment of elite mate search makes sense only if the market composition is grounded in public macro data and the outcome is framed as a scarce, capacity-constrained matching problem rather than a one-line classifier. In the resulting base case, target-match probabilities are low because the target subgroup is genuinely rare. The large differences arise not from age alone but from differential access to the overlapping communities in which elite urban partners are actually reachable.

The satirical version of the conclusion is simple: romance may resist optimization, but it still respects supply, queues, and network topology.

References

1. **Jaewon Yang and Jure Leskovec. "Overlapping Community Detection at Scale: A Nonnegative Matrix Factorization Approach (BIGCLAM)." WSDM 2013. PDF at Stanford.** https://cs.stanford.edu/people/jure/pubs/bigclam-wsdm13.pdf
2. **北京市统计局. "北京市2024年国民经济和社会发展统计公报." 2025-03-19.** https://tjj.beijing.gov.cn/tjsj_31433/tjgb_31445/ndgb_31446/202503/t20250319_4038820.html
3. **上海市统计局. "2024年上海市国民经济和社会发展统计公报." 2025-03-24.** https://tjj.sh.gov.cn/tjgb/20250324/a7fe18c6d5c24d66bfca89c5bb4cdcfb.html

4. **深圳市统计局. "深圳市2024年国民经济和社会发展统计公报." 2025-05-22 / 2025-06 official publication.**
https://www.sz.gov.cn/zfgb/2025/gb1374/content/post_12212437.html
5. **北京市统计局. "北京市第七次全国人口普查主要数据情况." 2021-05-19.**
https://tjj.beijing.gov.cn/zt/bjsdqcqgrkpc/qrpbjjd/202105/t20210519_2392982.html
6. **上海市统计局. "上海市第七次全国人口普查主要数据发布." 2021-05-18.**
https://tjj.sh.gov.cn/tjxw/20210517/4254aba799c840d2a54f9ef82858bcf5.html
7. **上海市统计局. "上海市第七次全国人口普查主要数据结果新闻发布会答记者问." 2021-05-19.**
https://tjj.sh.gov.cn/tjxw/20210519/529f29f128864bb8b31d74e4db9291ce.html
8. **深圳市统计局. "深圳市第七次全国人口普查公报." 2021-05-17.**
https://www.sz.gov.cn/zfgb/2021/gb1199/content/post_8806392.html
9. **上海市人民政府 / 市民政局. "2024年上海婚姻登记数据新鲜出炉！" 2025-01-24.**
https://www.shanghai.gov.cn/nw31406/20250126/a16ab149af98413b8bc3bfc21b87a2c0.html
10. **Yang, Jing. "China's Family Changes from the 2020 Population Census." East Asian Institute, National University of Singapore, 2023.**
https://research.nus.edu.sg/eai/wp-content/uploads/2023/03/EAIBB-No.-1696-Chinas-Family-Changes-2.pdf
11. **北京市人力资源和社会保障局. "历年北京市全口径城镇单位就业人员平均工资." 2025-09-18.**
https://rsj.beijing.gov.cn/bm/ywml/202007/t20200717_1950961.html
12. **上海市人力资源和社会保障局. "本市调整2025年度社保缴费基数上下限." 2025-09-18.**
https://rsj.sh.gov.cn/tgsgg_17341/20250918/t0035_1435637.html
13. **深圳市统计局. "2024年深圳市城镇单位就业人员年平均工资情况." 2025-06-30.**
https://tjj.sz.gov.cn/gkmlpt/content/12/12253/post_12253352.html
14. **国家统计局. "2024年城镇单位就业人员年平均工资情况." 2025-05-16.**
https://www.stats.gov.cn/sj/zxfb/202505/t20250516_1959826.html
15. **深圳市人力资源和社会保障局. "2024年人力资源市场工资指导价位——分行业工资价位表." PDF.**
https://hrss.sz.gov.cn/attachment/1/1572/1572960/12138162.pdf
16. **教育部. "第二轮'双一流'建设高校及建设学科名单." 2022-02-11/14.**
https://www.moe.gov.cn/srcsite/A22/s7065/202202/t20220211_598710.html
17. **教育部. "'211工程'与'985工程'." 2015-11-06.**
https://www.moe.gov.cn/jyb_xwfb/xw_zt/moe_357/jyzt_2015nztzl/2015_zt15/15zt15_mtbd/201511/t20151106_217950.html