Σ Cubing: Tireless Tracker - Analyzing Your Own Cube Drafts

Introduction

I've always loved data and statistics in the context of Magic (Frank Karsten is an idol of mine). When I fell in love with cube about four years ago, I was acutely interested in what data-based resources existed for cube design. But cube is a subjective enterprise. Different cubes have different goals in gameplay and drafting, and drafters have preferred playstyles and skill levels that can shift a cube metagame. Couple this with Magic's inherent complexity, and you have a recipe for disagreements in both card and archetype evaluation.

The goal of **Σ Cubing** is to answer questions about cube using real world data and to lay a solid statistical foundation for cube discussion. This is an ambitious goal, as data is never truly objective and riddled with biases, but you can get reasonably far with a well defined question and dataset. I've been collecting data on my own cube for almost 3 years, and I currently help the XMage Cube Group curate and analyze their 3-0 decklist dataset. I've analyzed cards like Experimental Frenzy and Demonlord Belzenlock to evaluate their power level through simulation. If a cube problem exists and can be looked at with data or simulation, I'll always give it a go.

Each installment of **Σ Cubing** will have at least four components.

- **Problem**: This section will discuss what aspect of cube we're trying to investigate. I'll discuss the problem's features and complexities and what answering the question might tell us about cube.
- **Data**: In this section, we'll discuss the dataset. I'll explain where the data comes from and how it was analyzed. I'll also cover any biases in the dataset.
- **Discussion/Results**: This is where I'll discuss my analysis of the dataset, what conclusions we can draw, and how confident we are in these conclusions.
- **Resources**: This section will contain links to the dataset, any code I used to analyze it, and any other resources pertaining to the article.

Tireless Tracker - Analyzing Your Own Cube Drafts

I talk often with people on the MTG Cube Talk Discord about tracking my own cube data, and I've been delighted to see that many also do this or are interested in doing so. The major hurdle everyone faces is finding a way to keep track of cube data in a way that is both efficient and fruitful. Keeping track of everyone's P1 and P15 is easy, but may not tell you much. Keeping track of every pick and every decklist's winrates is too time-consuming, even if it contains lots of useful data.

Problem

How should a cube owner collect data? I've been using this method for about two years now: after every draft, I ask the drafters to take a picture of their deck and send it to me. Later, I transcribe the decklists into a simple text file on my computer, noting the colors of the deck, its archetype, its record, and the cards it contains. I then use Python to parse all these files and combine their information into one dataset. That's all there is. I will provide the Python script and instructions for use in **Resources**.

Currently, the script outputs a few different analyses:

- **Archetype Analysis**: The script will analyze the win rates of each archetype and which cards appear most frequently in that archetype. It allows for subarchetypes – a UB Control Reanimator deck can be classified as both Control and Reanimator.
- **Card Analysis**: The script will analyze which cards appear most in your decklists and their archetype breakdown. It will also output individual win rates for each card (this feature comes with *significant* limitations in interpretability, see **Results**).
- **Color Balance**: It will analyze what colors are most often drafted in your cube. It will do this based both on the decks themselves and the cards in the decks. For example, if I have one UB Control deck in my dataset, blue and black share an equal archetype representation (0.5-0.5). But if that deck is playing 5 black cards and 17 blue cards, black will have a 5/23 = 0.217 *card* representation.

The script obtains card information (cmc, type, and color) from scryfall, so it currently needs an internet connection. With that, let's take a look at the data that was collected from my own cube using this script!

Data

Over the past 2-3 years, I've collected 399 decklists from drafts of my cube. I run a Strix Scale 8-F unpowered cube. In terms of design goals, I aim to maximize power and efficiency within the unpowered design restriction (for example, I cube Mana Drain and Mind Twist). Reanimator and creature-cheat strategies are well supported. I typically draft with 5-6 friends, occasionally a full 8 man draft or a 2-3 man Winston draft.

Biases

A bias is a trend in the data that exists as a result of some external force. In the case of drafted decklists, the primary bias is drafter preference. As a drafter, I love drafting aggressive decks, and will actively pick Goblin Guide/Sulfuric Vortex over most cards in the cube. Because I am the most experienced player in my playgroup at drafting cube, this means that aggressive decks may be overrepresented in terms of win rate.

A similar bias exists for individual cards. If skillful players think that mediocre a card is good, that card may have a high win rate because those players win more often. The same is true in reverse – good cards that skillful players underrate will have lower win rates than they should.

Results/Discussion

Archetype and Subarchetype Breakdown

In my own cube tracking, I've chosen to keep a higher order archetype breakdown (Aggro, Midrange, Control), and a sub-archetype breakdown (Ramp, Combo, Reanimator). This means that all decks are either Aggro, Midrange, or Control, but some have subarchetypes (Control-Reanimator, Midrange-Ramp, etc). Here are their winrates in my cube:

Archetype	Decks	Game Record	Win Rate
Aggro	102	445–328	0.58 ± 0.03
Midrange	169	656–658	0.50 ± 0.03
Control	135	490–532	0.48 ± 0.03
Ramp	63	300–214	0.58 ± 0.04
Combo	31	122-116	0.57 ± 0.06
Reanimator	25	74–97	0.43 ± 0.07

Table 1: Archetype Breakdown. ± 95% confidence interval calculated via boostrap.

Based on winrates, aggro and ramp decks are the top dogs in my cube. Seeing this data has encouraged me to introduce more tools against these archetypes into my cube (Pyroclasm, Whip Flare, Plague Engineer, etc).

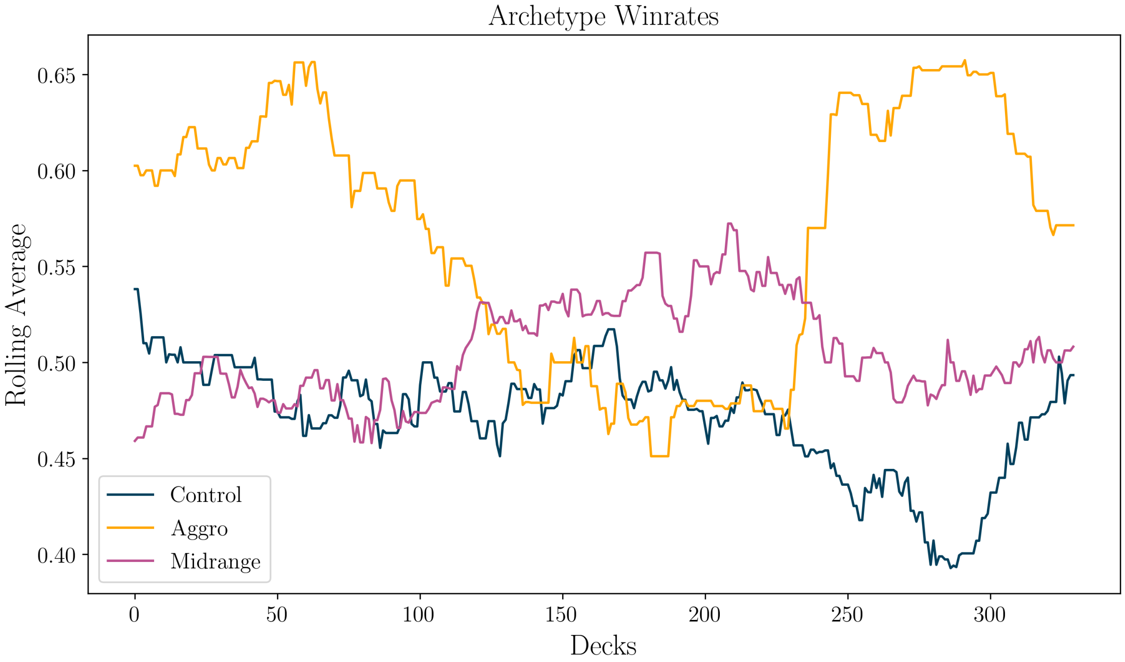
We can also investigate the common cards in each archetype:

Archetype	Most Common Cards
Aggro	Strip Mine, Sulfuric Vortex, Porcelain Legionnaire
Midrange	Polluted Delta, Recurring Nightmare, Demonic Tutor
Control	Ponder, Coldsteel Heart, Azorius Signet
Ramp	Birds of Paradise, Craterhoof Behemoth, Gaea's Cradle
Combo	Sneak Attack, Emrakul, the Aeon's Torn, Oath of Druids
Reanimator	Entomb, Reanimate, Griselbrand

Table 2: Most Common Cards in Each Archetype

It's clear that the most common cards in each archetype tend to either be archetype enablers (Sneak Attack, Reanimate), powerhouses in the archetype (Sulfuric Vortex, Craterhoof Behemoth), or flexible cards that will fit any color deck in the archetype (Strip Mine, Coldsteel Heart). This makes sense given that these cards either pull you into an archetype or are flexible in terms of color commitment.

Because the decklists are tagged with dates, we can also interrogate the change in these archetype winrates over time. To do this, I use a *rolling average*, which examines the average winrates of archetypes in a certain window of time. This enables us to make comparisons between time frames. In this case I use a 70 deck rolling average.



We can see from this graph that aggro has always been a powerhouse in my cube, although there have been times where it was not a top performer. I've taken a look at decklists during this time frame, and I've discovered that this was when the number of aggro decks per draft increased. The natural conclusion is that aggro's average winrate decreased because people fought for the archetype. I speculate that this is why aggro does so well in my cube generally – there is usually only one or two people drafting it. In theory, archetype win rates are self-correcting; players will realize which archetypes are the best and will compete to draft them, lowering their average win rate. Aggro decks likely dodge this self-correction, as many players who cube simply do not like playing aggro even if it is "optimal". As a cube designer, this presents a conundrum. Should I accept that the high win rates of aggro are a result of player preferences and do nothing to correct it? Do I provide tools to other archetypes against aggro, or does this "punish" drafters for recognizing that aggro is underestimated and drafting it? These are questions I haven't yet answered for myself as a cube designer.

Individual Card Winrates

When I first started collecting data on my cube, I hoped to evaluate the strength of individual cards. In theory, strong cards lead to strong decks, so maybe looking at the cards in winning decks could identify the performers and the duds. The easiest approach is to look at card "winrates". For example, if possessing a Tinker in your maindeck causes you to win every game you play, then the "winrate" of Tinker is 100%. I analyzed the cards that cards that have been in **more than 20 decks** (250 cards). I've chosen some illuminating examples:

Rank (Out of 250)	Card	Games	Win Rate
1	Fireblast	192	0.646
2	Figure of Destiny	188	0.628
3	Hexdrinker	224	0.620
5	Jackal Pup	197	0.614
10	Fyndhorn Elves	298	0.604
⋮	⋮	⋮	⋮
142	Mind Twist	197	0.501
173	Mana Drain	310	0.484
⋮	⋮	⋮	⋮
236	Vampiric Tutor	282	0.433
249	Toxic Deluge	268	0.411

Table 3: Individual Card Breakdown

You might notice that the individual card win rates do not have any associated uncertainties. This can make it difficult to interpret how confident we are in these win rates. When we calculate uncertainties, we make implicit assumptions about underlying properties of the data. For example, you may have seen standard deviation used to quantify uncertainty. This makes strong assumptions about how the data were generated (namely, that the data are from a normal distribution). Simulation is a much better way to explore uncertainty in this case, but first let's cover the most obvious thing about this analysis:

The "win rates" of individual cards are entirely reflective of archetype win rates, not individual card strength.

Mind Twist and Mana Drain are probably the strongest cards in my cube, yet their winrates pale in comparison to the mighty Jackal Pup. This is simply because aggro decks are dominant in my cube, and those decks *happen* to often contain Jackal Pup. Same for the ramp cards like Hexdrinker and Fyndhorn Elves. Conversely, Vampiric Tutor and Toxic Deluge aren't bad cards; they just happen to go in decks that do relatively poorly in my cube compared to aggro.

"Now hold on!" you might say. *"That wouldn't be an issue if your cube wasn't an aggro slugfest. If your archetypes were balanced, we would see cards sort by relative strength."*

Well, yes and no. It is true that if my archetypes were balanced, these win rates would be *more* reflective of individual card strength. But it would still be immensely difficult to estimate card strength accurately. This is because predicting a card's strength based on the raw winrates of decks that contain it only works with an insane amount of data

Simply put, we don't have enough decklists to accurately estimate individual card strength. It is a common problem in data science to estimate parameters from input data. In our case, we're trying estimate 450 parameters (card strengths) with only about 400 decklists and their winrates (input data). This is like trying to predict an NBA player's three point percentage by watching them play for 15 minutes. This begs the question of how many decklists we would need to accurately estimate card strength, a question we can answer with simulation.

In statistics, simulation is the process of modelling a process with a computer to learn more about its properties. Every simulation needs a **generative model**, or a sequence of steps that produces data. We can devise a generative model for how card win rates are produced:

- We'll have a 450 card cube, with each card having a number indicating its "strength" from 0-10. Most cards have a strength around 5*.
- To make a deck, we randomly choose 45 cards from the cube, then choose the best 23 of those cards. The deck's total strength is the sum of the strengths of its cards.
- Decks play against each other in best of three. For a given game, a deck has a probability to win equal to its strength divided by the total strength of the two decks. For example, if a deck with strength 100 plays against one with strength 150, it has a 100/250 = 40% chance of winning a game.
- In each "tournament", eight decks will play against each other with three total matches.

To evaluate how many games we need to see strong cards have high win rates, we can devise a test. Let's remove 20 random cards and replace them with cards that have strength 10 (better than 99% of the cube). This is known as a "spike-in". We can test how many games it takes to see these spike-in cards develop higher win rates than other cards.

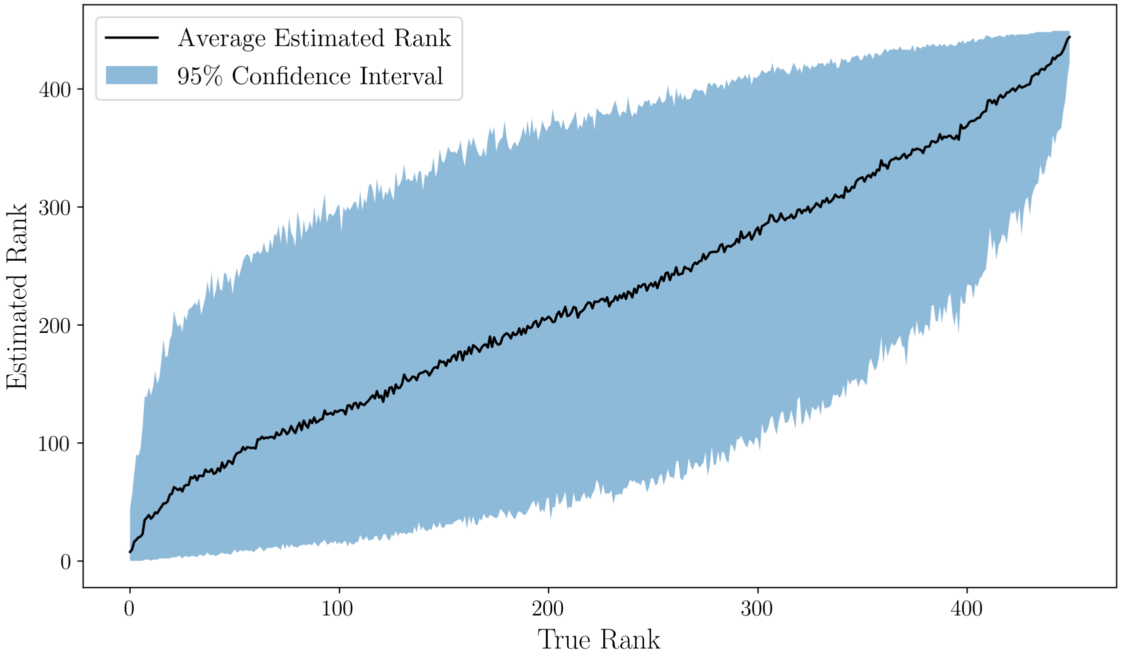
We can run a set number of tournaments, and identify both the average rank of the spike-in cards and the percentage of spike-in that ended up in the top 100 win rates:

Number of Tournaments	**Average # of Games/Card**	**Average Spike-in Rank (out of 450)**	**% of Spike-in in Top 100**
0	0	225	22%
100	300	124	43%
500	1,500	99	57%
1,000	3,000	83	66%
10,000	30,000	32	99%
∞	∞	10	100%

Table 4: Individual Card Spike-in Simulation. 0 tournaments played refers to having no information (ie if each card were randomly assigned a rank). ∞ tournaments refers to having perfect information on the cards. These are our baselines of reference.

We can see that only when approach the 30,000 game treshold do we start to see fully accurate ranking of the spike-in cards. Furthermore, this model is an extreme simplification

*Specifically, the card strengths will be normally distributed with mean 5, standard deviation 2 of cube; it does not account for the fact that sometimes cards are not drawn in games and cards can vary in strength between games. These additional sources of variation and bias will make accurate ranking of cards more difficult. It is also trying to detect *very* strong cards (10s on a scale of 0-10), since these cards separate out from others the fastest.



The plot above shows our confidence in the estimated rank of cards based on their true rank. To make this plot, I ran a 10,000 tournament simulation, noting how the cards ranked based on win rate (estimated rank) against their true rank. I then repeated this process 20,000 times to determine the average estimated rank and the confidence interval around this value. The graph shows that we can be very confident in strong cards (rank ≈ 1) and weak cards (rank ≈450) after 10,000 tournaments, as indicated by the narrow confidence interval. But for cards that are in the middle (rank ≈ 200), the confidence interval is huge. This means that if you estimated that a card is ranked 225th based on a 10,000 tournament dataset, the card is actually likely to be truly ranked anywhere from 30th to 400th.

I estimate that you would need upwards of 100,000 games to evaluate strong or very weak cards in *real* cubes and probably millions to evaluate average ones. So despite my hopes, it's basically impossible to evaluate the strength of individual cards with just raw win rates.

But not all hope is lost. By looking at just the win rates, we ignore other important variables (most notably, other cards present in the decks) and hope they just average out. There are statistical frameworks that can account for these variables. We could, for example, fit our data to the generative model described above. This approach, known as *hierarchical modeling*, allows us to account for all the cards in a deck and would estimate the actual strength of cards on a 0–10 scale. We'll explore this method in a future installment.

Color Analysis

The script will also analyze the color distribution of the decks. Looking at my dataset,

Color	Deck %	Average Card %
W	38%	44%
U	60%	37%
B	42%	40%
R	40%	44%
G	31%	56%

Table 5: Color Analysis. Deck % refers to the percentage of decks that contain the color. Card % refers to what fraction of cards are that color in a deck that contains that color.

Clearly, blue is the top dog in my cube in terms of what cards are drafted. But while 60% of decks play some amount of blue, those tend to play fewer blue cards relative to their other color. This implies that blue is being fought over and is often splashed. Conversely, decks playing green play 56% green cards, which means that green decks tend to be heavy green. This makes sense, given that green is not a color you splash often.

Conclusions

This analysis toolkit is clearly most helpful in tracking archetype win rates and balance in your cube. Before doing this analysis, I had assumed that Ramp was underperforming based on my personal experience; I never seemed to do well with MonoG decks or G/X ramp decks. As a result, I introduced changes several months ago that were designed to boost ramp decks. Those changes caused Ramp deck win rates to sky rocket; in one time frame, they were pushing 70%. Upon looking at the data, I was surprised to find that Ramp has always done well, and my changes turned a good archetype into one that was upsetting the balance of my cube. Ramp's win rate is gradually dropping as I introduce tools against it and the metagame self-corrects, but this serves as a warning against making sweeping changes based on human perception.

The individual card analysis was not as fruitful as I'd hoped. We can't rank cards based on win rates alone, and the win rates are mostly reflective of archetype balance anyway. But there are still interesting things to learn about individual cards. I've begun collecting sideboard information for my decklists; I'm very interested in seeing which cards are often sideboarded in decks that could otherwise play them.

Resources

Insert resources here.

Up Next

In our next installment, we'll be analyzing over 100,000 decklists from CubeTutor!